University of Sussex
Falmer
Brighton
2019

UNIVERSITY OF SUSSEX

Exploring the history of star formation in galaxies and its environmental dependence at high redshift

Christopher C. Lovell

Submitted for the degree of Doctor of Philosophy

University of Sussex

Supervisors: Peter A. Thomas & Stephen M. Wilkins

Abstract

The first stars formed in the early universe and shortly after assembled into the first galaxies. Since then, galaxies have been subject to a variety of processes, both internal and external, that affect their ability to form stars. At low redshift, environment plays a large role in inhibiting star formation, however it is less clear what effect it has at high redshift. This is predominantly due to the difficulty of determining the nature of the high redshift environment from uncertain redshift measurements, and the small coverage of high redshift surveys leading to poor sampling of the cosmic variance.

In this thesis I use a variety of numerical approaches to various aspects of this problem. In the first section I use a semi-analytic model to study the relationship between observed galaxy surface overdensity and the probability of coinciding with a protocluster, the pre-collapse progenitors of galaxy clusters, and make recommendations for optimum measurement apertures for their identification. In the second section I use a suite of hydrodynamic simulations of galaxy clusters, across a range of descendant halo masses, to study the galaxy evolution in their protocluster progenitors in detail. I characterise the star-forming sequence, studying it's difference in protocluster and field environments, as well as within dense groups in the collapsing protocluster.

In the final section I use a novel approach to estimate the star formation history of galaxies. Rather than studying the high redshift environment directly, I estimate when the stars in a low redshift galaxy were formed using population synthesis techniques. In this work I couple this with hydrodynamical simulations in order to provide more informative priors on the shape of the star formation history, which typically imposes strong biases on inferred properties, such as the total stellar mass, in more traditional approaches.

Keywords – Galaxy Protoclusters, High-z Galaxy Evolution, Machine Learning

Declaration

I declare that no part of this work is being submitted concurrently, or has been submitted previously, for another award of the University or any other awarding body or institution.

The following parts of this submission have been published previously:

- Chapter 3: 'Characterising and Identifying Galaxy Protoclusters', this has been published as Lovell et al. 2018, Monthly Notices of the Royal Astronomical Society, 474, Issue 4, p.4612-4628
- Chapter 5: 'Learning the Relationship between Galaxies Spectra and their Star Formation Histories using Convolutional Neural Networks and Cosmological Simulations', accepted for publication in Monthly Notices of the Royal Astronomical Society, 12$^{\text{th}}$ October 2019, https://doi.org/10.1093/mnras/stz2851

The following chapters are in preparation for submission:

- Chapter 4: 'Galaxy Protoclusters in the Cluster-Eagle project: evolution of the star-forming sequence', *in prep.*

The text of these papers has been modified to better fit with the rest of the thesis.

University of Sussex

May 2019

Acknowledgements

It's been a long and particularly bumpy road to get to this point, but a journey that I have enjoyed immensely, which is all thanks to the people I got to share it with. First, I'd like to thank my supervisors, Peter and Steve, who have both been incredibly patient and supportive from the very first day. I have learnt so much from you both, particularly on the subject of acronym creation. A huge thank you to Viviana, who hosted me in the Big Apple. You are the most amazing teacher and mentor, doing absolutely incredible things with your students, and I find your drive and energy so inspiring. Shout out to the rest of the CUNY crew, keep it punny. A big thanks to Romeel, who hosted me in Cape Town, where I met the incredible Kate, who isn't bad for a cosmologist. Also thanks are due to David for his infinite patience getting me up and running with the EAGLE code. Thank are due as well to my viva examiners, Dr. Robert Smith and Prof. Nina Hatch, for a stimulating and unexpectedly enjoyable discussion, and to the University of Sussex and the Science and Technologies Facilities Council (STFC) for their support.

To all the wonderful people in the Astronomy Centre, old and new. To name a few, Scott, Benoit, David, Steve, Rose, Carlos, Azizah, Jess, Ciaran, Pippa, Sunayana, Dimitrios, Aswin, Ian, Jussi, Luke, Kareem, and of course the rest of the Falmer Friday crew. My liver will be glad to see the back of you. To those that have been there for the highs and the lows, Annie, Mark, Dylan, Rory, Richard and Lauren. Finally, I'd like to thank the Kernow gang. My sister, who inspires me more than she knows. And Mum & Dad, for instilling in me a love of learning throughout my life, and without whose support (emotional and financial!) I would not be here today.

"It is good to have an end to journey toward;
but it is the journey that matters, in the end."

URSULA K. LE GUIN

Contents

List of Figures

List of Tables

1 Introduction

In this thesis I explore the history of star formation in galaxies and its environmental dependence at high redshift. Below I set out the main scientific themes of each chapter, and a summary of the results.

In Chapter 3 I use a Semi-Analytic model to explore methods of identifying and characterising galaxy protoclusters more robustly. I find that a significant fraction of all galaxies reside in protoclusters at $z > 2$, particularly the most massive, motivating their detailed study in this thesis. I pay particular attention to their spatial distribution, testing for the best aperture to measure the galaxy overdensity in to obtain a good correlation between the measured overdensity and the descendant mass (a useful protocluster diagnostic), maximising the completeness and purity of the galaxy population, and find indirect evidence for the emergence of a passive sequence in protoclusters at $z \sim 2$. I also present the first characterisation of protocluster shapes as traced by their galaxy populations, showing that they tend to be aspherical with a prolate distribution. The relationship between AGN and protoclusters is also investigated. Finally, I present a new procedure for estimating the probability that a given overdensity represents a true protocluster, and provide relations to estimate the descendant cluster mass.

In Chapter 4 I study the properties of protocluster galaxies in greater detail, utilising detailed hydrodynamic zoom simulations from the C-Eagle project. In particular I study the star-forming sequence in both protoclusters and the field to determine whether protocluster environment has any effect, at fixed stellar mass, on a galaxies SFR at high redshift. I also explore the scatter in the star forming sequence, which contains information on short timescale fluctuations in the star formation history, and compare to a number of recent observational constraints at high redshift. Finally, I study the passive galaxy population, and show how the passive fraction appears to be environmentally independent in the C-Eagle model, but shows significant environmental dependence in protocluster observations at $z \sim 2$.

Finally, in Chapter 3 I investigate an alternative way of studying the history of star formation, by analysing the integrated spectral energy distribution (SED) of individual, present day galaxies ($z \sim 0$). I present a new approach, combining the outputs of

cosmological simulations, coupled with detailed spectral energy distribution modelling, to train a supervised regression machine learning method. This provides a means of using the self-consistent information from the simulation on the SFH and its interdependence with the observed SED to extract better priors on the SFH from observations.

2 Background

2.1 Cosmological Background

The standard cosmological model, known as the Lambda Cold Dark Matter (ΛCDM) model, predicts the formation and evolution of structure in the Universe (Peebles, 1984). Its main components are baryonic matter, cold dark matter (CDM) and dark energy (Λ), the latter being responsible for the accelerating expansion of the universe (Peebles, 1993). In this model, during a period of inflation which began approximately 10^{-36} seconds after the Big Bang, the universe expanded by a factor of 10^{26} and structure was seeded by quantum fluctuations, which were blown up during inflation and evolved over time into inhomogeneities on a range of scales (Guth, 1981; Liddle & Lyth, 2000). Over the next ~ 20 minutes of cosmic time after the Big Bang, primordial nucleosynthesis determined the abundances of the light primordial elements (^{4}He, D, ^{3}He and ^{7}Li, Coc & Vangioni, 2017).

Approximately 370 000 years later, neutral atoms began to form from the ionised plasma. This episode of 'recombination' led to decoupling of photons from the rest of the baryonic matter, increasing their mean free path until they travelled essentially uninhibited to the present day; due to cosmological redshift they are now observed at millimetre wavelengths as the CMB (Penzias & Wilson, 1965; Planck Collaboration et al., 2014). In the absence of photon pressure the baryonic matter could collapse into the potential wells already formed by the non-interacting CDM, and continue to collapse through radiative processes, whereas the dark matter, without such cooling channels, remained in an extended, diffuse halo. The Universe remained neutral for the next half a billion years, a period known as the 'cosmic dark ages', until the first stars and galaxies formed in the collapsed halos. These first objects not only lit up the universe after the dark ages, but also released UV photons that began the process of cosmic reionisation. It is this 'stelliferous' period of the Universe's history with which this thesis is concerned.

The combination of measurements of the CMB at $z \sim 1100$ as well as of the large scale structure traced by galaxies in the lower redshift universe, provide stringent constraints on the ΛCDM parameters (Tegmark et al., 2004; Planck Collaboration et al., 2014).

Unless otherwise stated, I assume a Planck 2013 cosmology throughout this thesis (Planck Collaboration et al., 2014), with the following parameters: $\Omega_m = 0.30$, $\Omega_\Lambda = 0.69$, $\Omega_b = 0.048$, and $h = 0.68$. Here, Ω_m describes the total matter fraction, Ω_Λ the dark energy fraction, Ω_b the baryonic matter fraction, and h the hubble parameter, assuming the Universe is close to the critical density, $\Omega = 1$.

2.2 Astrophysical Background

2.2.1 Formation of the First Stars and Galaxies

Structure in ΛCDM forms hierarchically through the merger of dark matter halos, forming progressively more massive structures with decreasing redshift. The first stars are thought to have formed between $30 > z > 20$ in dark matter 'minihalos' with mass $\sim 10^6 M_\odot$ (Bromm & Yoshida, 2011; Greif, 2014). The primordial abundance of the gas involved in the formation of the first stars is the source of the name 'Population III', leading on from the Population I and II stars in the local universe defined by their (relative) metal richness or metal deficiency, respectively (Bromm & Larson, 2004; Glover, 2013). Early studies suggested that the Initial Mass Function (IMF) of Pop.III stars was 'top-heavy', with greater relative numbers of very high-mass stars, due to the absence of metals or dust (Larson, 1998; Greif et al., 2011; Bromm, 2013). This left H_2 as the only cooling channel; with a temperature floor of ~ 300 K, this gives a characteristic Jeans mass of $\sim 10^3\ M_\odot$ (Benson, 2010).

At the ends of their lives some fraction of the Pop.III stars (dependent on the assumed IMF) are expected to go supernovae (SNe). These SNe distributed the heavier chemical elements formed within, known as 'metals' in Astronomy phraseology, into the pristine primordial Inter-Galactic Medium (IGM). This metal-enriched IGM enabled more efficient subsequent cooling channels for the gas (Rees & Ostriker, 1977; Silk & Wyse, 1993), leading to a rapid acceleration of subsequent star formation (Population II) and the build up of the first galaxies. The first generation of stars also produced dust, which provided new channels for H_2 formation, enabling more cooling.

The collapse of the first stars led to the first stellar black holes (as distinct from Primordial Black Holes, Carr & Hawking, 1974). These are one potential source for observed high-z

supermassive black holes (SMBHs). These high-z SMBHs are problematic, since they require either very massive seeds, or prolonged super-Eddington accretion in order to reach their observed masses at $z \sim 7 - 6$. Other formation mechanisms for these high-z SMBHs are Direct Collapse Black Holes (DCBHs) formed from the monolithic collapse of primordial gas in atomic cooling halos (Smith & Bromm, 2019).

2.2.2 Reionisation

The formation of the first stars and galaxies released UV photons into the IGM, ionising the neutral hydrogen and kick-starting the Epoch of Reionisation (EoR). At the highest redshifts the ionisation was limited to the immediate surroundings of these collapsed objects, creating a 'swiss-cheese' topology of ionised bubbles and neutral regions between them (Zaroubi, 2012). Not until a sufficiently large number of galaxies formed throughout the cosmic web did the volume filling fraction of ionised hydrogen reach significant levels, until at some point the majority of the Universe was ionised, signalling the end of the EoR.

The timing of the EoR has been constrained by a number of measurements. The Gunn-Peterson trough in Quasar spectra (Gunn & Peterson, 1965) suggests an increasingly neutral IGM at $z > 6$ (Becker et al., 2001), placing a lower limit on the end of the EoR. Another constraint is provided by the optical depth to Thomson scattering of the CMB, which gives an instantaneous reionisation redshift of $z = 7.68$ in the latest Planck results (Collaboration et al., 2018). In future, the Square Kilometre Array (SKA) will map the topology of reionisation by observing the redshifted 21-cm emission from neutral hydrogen, distinguishing the bubbles of ionised material and their spatial evolution throughout the EoR (Mellema et al., 2013; Datta et al., 2016; Trombetti & Burigana, 2018).

The sources of the ionising photons are still unknown, though the currently favoured candidates are star-forming low-mass galaxies; observations with the Hubble Space Telescope (HST) suggest a steep low-mass end of the UV Luminosity Function (UVLF), providing sufficiently large numbers of ionising photons (Bouwens et al., 2012; Ellis et al., 2013). Low-mass galaxies are also expected to have a higher escape fraction of ionising photons, $f_{\rm esc}$, since strong stellar feedback can create channels in the ISM through which ionising photons can escape (Wise et al., 2014). However, the inclusion of additional

physics, such as the effect of binary interactions, could increase $f_{\rm esc}$ for high-mass galaxies (Ma et al., 2016). Active Galactic Nuclei will also emit ionising radiation, and have been touted as a potentially significant contributor to the ionising photon budget (Madau & Haardt, 2015), though this has recently been considered less likely due to their low number densities at $z > 6$ (Hassan et al., 2017; Parsa et al., 2017).

The history of reionisation inferred from the UVLF now matches that inferred from CMB measurements (Robertson et al., 2015). Recent studies have suggested using the topology of reionisation itself to infer the shape of the UV Luminosity Function at high-z, assuming some mapping between the sources and their ionised bubbles (Zackrisson et al., 2019).

2.2.3 Stellar Populations

2.2.3.1 Star Formation

Star formation occurs in dense regions of a galaxy's Interstellar Medium (ISM) where molecular hydrogen (H_2) can form. When the internal gravitational potential energy of the cloud is greater than the outward pressure of the thermal kinetic energy it collapses; the mass at which this criterion is satisfied is known as the Jeans Mass (M_J) after Sir James Jeans (b.1877), and is parametrised as

$$M_J = \left(\frac{5k_bT}{Gm}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2} \quad , \tag{2.1}$$

where m is the mean particle mass, T is the average temperature, k_b is the Boltzmann constant, and G is the gravitational constant. $M_J \propto \rho^{-1/2}$, which leads to an interesting property: as the cloud collapses and the density increases, the jeans mass decreases, which causes hierarchical fragmentation of the cloud. Collapse can also be triggered by external events, such as shockwaves from nearby supernovae or collisions of clouds. Once the fragments reach sufficiently high densities and temperatures, nuclear fusion is initiated and the protostar joins the main sequence.

Stars form predominantly in binary or higher multiple systems, with a wide range of separation, eccentricity and mass ratio (Eldridge et al., 2017).

2.2.3.2 Initial Mass Function

The distribution of masses in a forming stellar cluster is known as the Initial Mass Function (IMF),

$$\frac{\mathrm{d}n}{\mathrm{d}M} = \phi_{\mathrm{IMF}}(M) \ . \tag{2.2}$$

Whilst not directly observable, the IMF can be empirically derived from observations of stellar clusters at later stage of their lifetimes, by counting the relative number of stars of different masses, and hence differing lifetimes. Local measurements in the Milky Way suggest a near universal form that is independent of star-forming conditions (Hopkins, 2018). It was first parametrised by Salpeter (1955) as a single power law,

$$\phi(M) = \beta\, M^{\,\alpha} \tag{2.3}$$

where β gives the normalisation, and the exponent $\alpha = -2.35$. This form predicts very high numbers of sub-solar mass stars that are hard to reconcile with most recent measurements of the stellar luminosity function. Later authors have implemented a broken power law form, with a shallower slope below $\sim 1\,M_{\odot}$ (Miller & Scalo, 1979; Kroupa, 2001; Chabrier, 2003).

Extragalactic measurements tentatively suggest a variable IMF, possibly proportional to the galaxies Star Formation Rate (SFR) (Lee et al., 2009; Gunawardhana et al., 2011; Zhang et al., 2018).

The lower mass limit of the IMF, at which the clump size is insufficient to achieve the high density and pressure to initiate fusion, is $\sim 0.08 M_{\odot}$. Lower-mass objects can collapse below this threshold, known as Brown Dwarfs, but their contribution to the total cluster mass is assumed to be low due to the steep turnover of the low-mass IMF (Kroupa et al., 2013).

2.2.3.3 Stellar Evolution

Stars spend the majority of their lives on the main sequence, regardless of their mass. The length of this phase is determined by the amount of core hydrogen and its rate

of consumption in the fusion process. Whilst high-mass stars have greater amounts of hydrogen fuel the rate of consumption is higher, so the main sequence lifetime is inversely proportional to the stellar mass.

The final stages of a star's lifetime is also highly mass sensitive. For intermediate age stars this proceeds through a red giant phase, with significant mass loss from the outer layers, leaving an inert core supported by electron degeneracy pressure, known as a white dwarf. More massive stars ($M_{\odot} > 40M_{\odot}$) exceed the Chandrasekhar limit (Chandrasekhar, 1931) and are unable to support their core through electron degeneracy pressure, leading to collapse down to a neutron star held up by neutron degeneracy pressure. This process involves significant energy release through a core collapse supernovae (CCSN); these events produce the majority of elements heavier than Iron, injecting them, as well as significant energy, into the ISM.

Even more massive objects will exceed the limit of neutron degeneracy pressure, and collapse to a stellar mass black hole.

2.2.4 Galaxy Demographics and their Evolution

Galaxies exhibit an incredible diversity among a number of key properties, both directly observed and intrinsic properties inferred through Spectral Energy Distribution (SED) fitting. A key aim of the field of galaxy evolution is to determine the cause of this diversity, and to reproduce it in models, both in individual objects and the relative abundances statistically.

A common approach is to derive intrinsic properties of observed galaxies through SED fitting and perform comparisons to models in this physical parameter space. The alternative is to *forward model* the simulations to produce observed distribution functions of *e.g.* UV luminosity, or emission line distribution functions. Each approach requires differing modelling techniques reliant on differing assumptions and biases. Using a combination of approaches helps to elucidate these biases.

The two galaxy distribution functions studied in detail in this thesis are the stellar mass function and the stellar mass-star formation rate distribution, also known as the main sequence or star-forming sequence. Below I describe each one, and the proposed physical mechanisms that shape them.

2.2.4.1 The Galaxy Stellar Mass Function

The Galaxy Stellar Mass Function (GSMF) describes the number of galaxies per unit volume per unit stellar mass interval $\mathrm{d}M$,

$$\phi(\mathrm{M}) = N \,/\, \mathrm{Mpc}^{-3}\,\mathrm{dex}^{-1} \quad , \tag{2.4}$$

and is commonly described using a Schechter function (Schechter, 1976),

$$\phi(M)\mathrm{d}M = \phi_* \left(\frac{M}{M_*}\right)^{\alpha} \exp\left(\frac{-M}{M_*}\right) \frac{\mathrm{d}M}{M_*} \quad , \tag{2.5}$$

which describes the high- and low-mass behaviour with an exponential and a power law dependence on stellar mass, respectively. Recent studies have found that a double Schechter function can better fit the full distribution (*e.g.* the GAMA survey, Baldry et al., 2008).

In the first cosmological models of galaxy evolution the stellar mass function was poorly reproduced at the high- and low-mass end, until the inclusion of two feedback mechanisms: stellar and AGN feedback. This behaviour can also be seen in the stellar to halo mass relation (SHMR), which links the stellar and halo mass functions. The SHMR has a peak at a halo mass of $\sim 10^{12} M_{\odot}$ and declines for low- and high-mass galaxies either side of the peak, with scatter of ~ 0.2 dex around this relation. At low stellar masses, energetic stellar feedback efficiently ejects gas from the halo, lowering the SFR (Kauffmann et al., 1993), and reducing the number density of low-mass galaxies. At the high-mass end, very high-mass SMBHs have been indirectly observed (the correlation between halo mass and SMBH mass is known as the Magorrian relation; Magorrian et al., 1998), that can maintain high accretion rates. This leads to efficient AGN feedback, which ejects gas from the central regions of massive galaxies, leading to a reduction in number density of high stellar mass galaxies (White & Rees, 1978). Modern Semi-Analytic Models are now able to reproduce the low- and high-mass behaviour of the GSMF with the inclusion of these two principal feedback processes (see Section 2.3.3).

Figure 2.1 shows the GSMF in the fiducial Eagle simulation (described in detail in Section 2.3.4.2) along with a number of observational constraints up to high redshift.

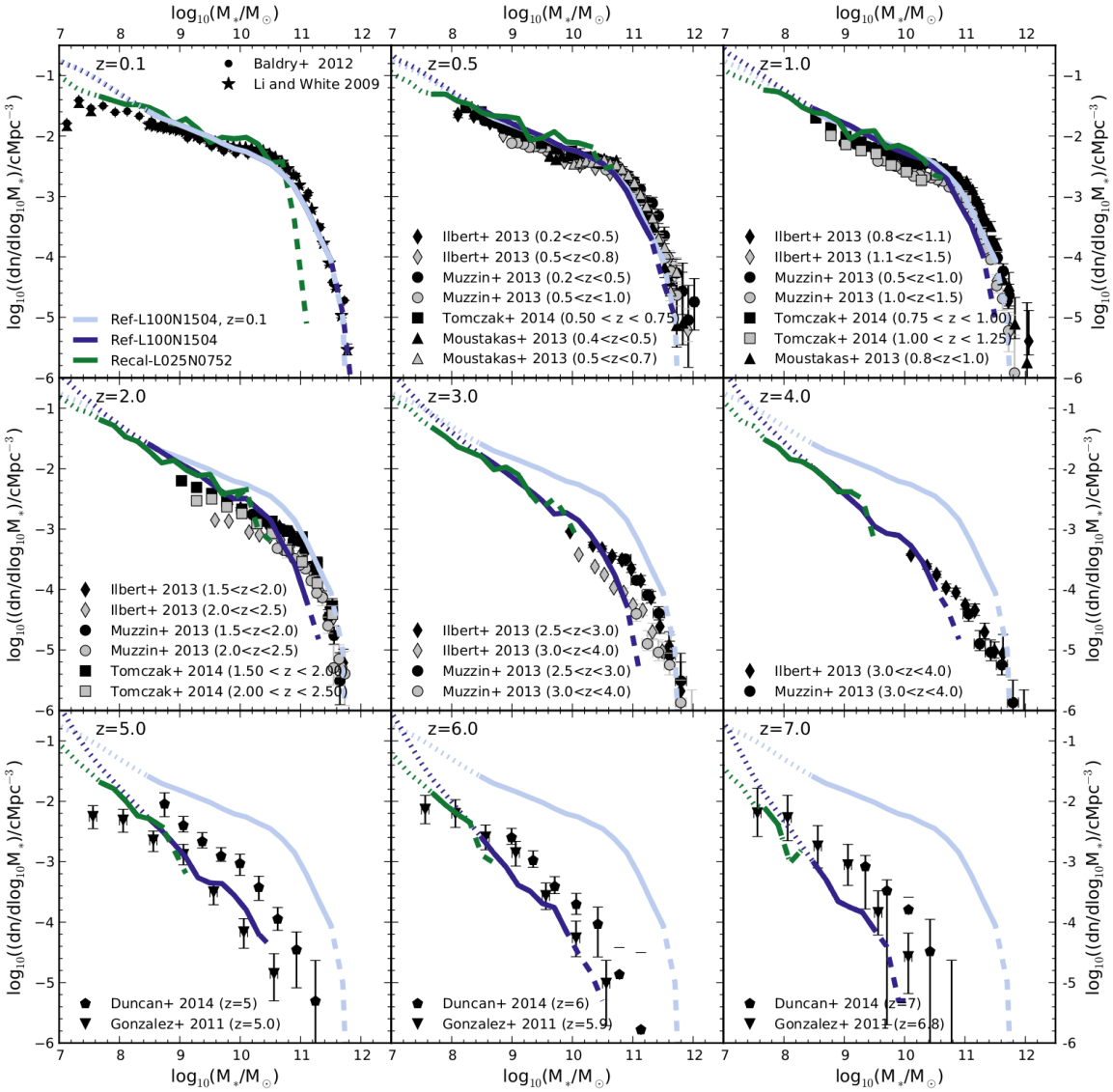

Figure 2.1: The evolution of the GSMF in the fiducial (Ref) and recalibrated (Recal) EAGLE simulations. Observational constraints are shown from Ilbert et al. (2013); Muzzin et al. (2013); Tomczak et al. (2014); Moustakas et al. (2013); Duncan et al. (2014); Gonzalez-Perez et al. (2014). Reproduced from Furlong et al. (2015).

2.2.4.2 The star-forming sequence

Observations suggest a close relationship between the star formation rate (SFR) and stellar mass of galaxies, at both high and low redshifts, which I will refer to as the star-forming sequence (SFS), though it is also commonly referred to as the 'main sequence' (Brinchmann et al., 2004; Noeske et al., 2007; Speagle et al., 2014). The SFS is typically parametrised as a linear relation,

$$\log_{10}(\mathrm{SFR}) = \alpha \log_{10}(\mathrm{M}_* \,/\, \mathrm{M}_\odot) + \beta \quad , \tag{2.6}$$

where the slope α remains relatively constant with increasing redshift, but the normalisation β increases (Daddi et al., 2007; Santini et al., 2009; Salmon et al., 2015). There have also been suggestions of a turnover in the SFS at high stellar masses (Lee et al., 2015; Tasca et al., 2015), though the turnover becomes less evident with increasing redshift. This behaviour has been seen in both low redshift (Lee et al., 2015) and high redshift (Tasca et al., 2015; Santini et al., 2017) observations, but is absent from some models (*e.g.* Illustris, Sparre et al., 2015). The turnover may be evidence for a change in the dominant channel of stellar mass growth from smooth gas accretion to merger driven growth. A high-mass SFS turnover is also necessary to explain the galaxy stellar mass function at lower stellar masses; a single power law slope would lead to too many massive galaxies being formed (Leja et al., 2015).

2.2.5 Galaxy Star Formation Histories

The instantaneous SFR of a galaxy at different times throughout its history is known as its Star Formation History (SFH). By definition, the integral of the SFH up to the observation time gives the total stellar mass at that time,

$$M_*(t_{\mathrm{obs}}) \,/\, M_\odot = \int_0^{t_{\mathrm{obs}}} (\mathrm{SFR}(t) \,/\, M_\odot \,\mathrm{yr}^{-1})\, R(t_{\mathrm{obs}} - t)\, \mathrm{d}t \quad , \tag{2.7}$$

where R is the recycling fraction for a stellar population born at time t, which accounts for the fact that stars return mass to the ISM at the end of their lives through *e.g.* supernovae, and this increases with time as a greater fraction of stars reach the end of their lives.

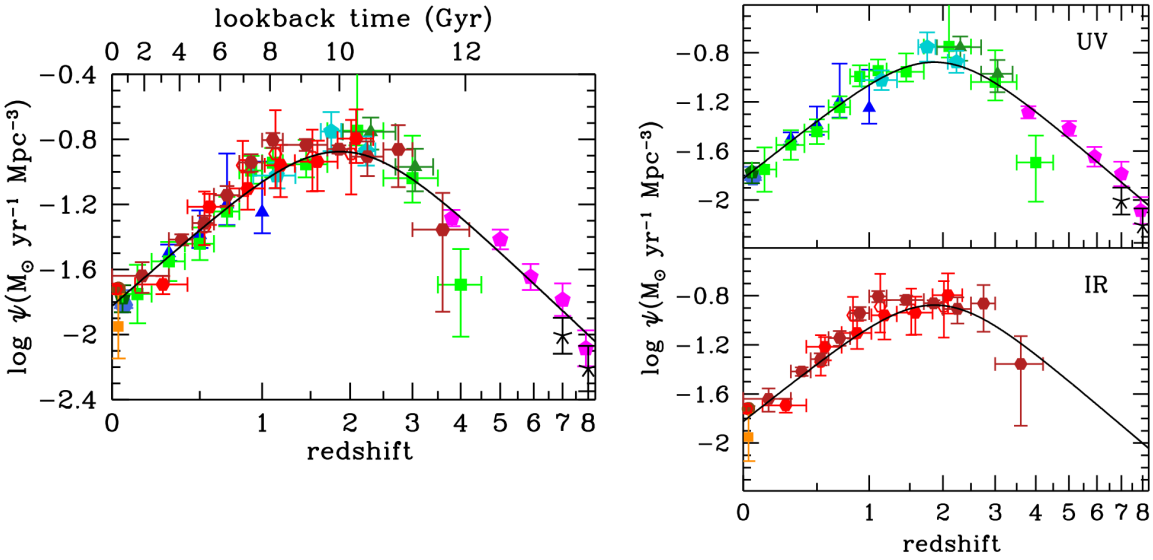

Figure 2.2: The observed cosmic star formation rate density as a function of redshift, assuming a Salpeter IMF (Salpeter, 1955), from both UV and IR tracers. The solid black curve shows the best fit to the observations. Reproduced from Madau & Dickinson (2014).

The SFH of all galaxies normalised by volume is known as the Cosmic Star Formation Rate Density (CSFRD), which shows a characteristic shape, rising from early times before peaking at cosmic noon ($z \sim 2$), then falling again towards the present day (see Figure 2.2). However, individual galaxy histories are diverse, showing bursting and quenching behaviour, as well as stochasticity on short timescales.

When fitting the Spectral Energy Distribution (SED) of a galaxy one typically assumes a parametric form for the SFH, that is then used to derive the total stellar mass and current SFR (Carnall et al., 2019). Common parametrisations of the SFH include exponentially declining, lognormal and double power law forms; Figure 2.3 shows a number of these compared to a 'true' SFH taken from a Semi-Analytic Model, fit to the same noisified SED (Somerville et al., 2008). Composite SFHs can also be described that combine many evolutionary features of galaxies seen in observations, such as rising, burst-like or quenching histories (Iyer & Gawiser, 2017). It has been shown that the choice of parametrisation can lead to significant biases in inferred properties (Acquaviva et al., 2011; Iyer & Gawiser, 2017; Carnall et al., 2019).

As well as parametric approaches, a number of non-parametric approaches have been proposed to account for a range of SFH behaviours (e.g. MOPED & VESPA, Heavens et al., 2000; Tojeiro et al., 2007, 2009). However, such approaches are highly sensitive to

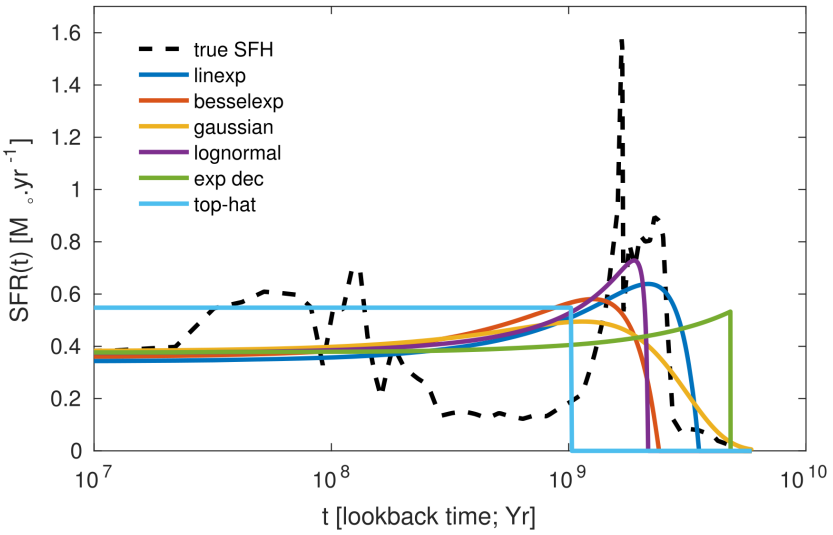

Figure 2.3: Common parametric forms for the Star Formation History (SFH). The black dashed line shows a SFH taken from a Semi-Analytic Model (Somerville et al., 2008), and each coloured line shows the best fit SFH to reproduce the noisified SED. Reproduced from Iyer & Gawiser (2017).

the chosen prior distribution (Leja et al., 2019). I will show in Chapter 5 how cosmological simulations can be used to provide an informative prior.

2.3 Cosmological Simulations

As we have seen so far, the physics of galaxy formation is a complex mix of processes operating over orders of magnitude in scale and energy. It is therefore a challenge to model these processes self-consistently, particularly over the large cosmological volumes required to obtain a large, representative sample of galaxies (Somerville & Davé, 2015). As a result, a number of approaches have been developed that model these processes with varying levels of sophistication, and resulting differences in computational complexity and cost. Below I describe the N-body method used for simulating collisionless fluids, such as CDM, and two methods used in this thesis for modelling the baryonic content, Semi-Analytic models (SAMs) and Hydrodynamic simulations. I also discuss common structure finding algorithms for both dark-matter only and hydro simulations.

2.3.1 N-body simulations

Cold Dark Matter (CDM) dominates the mass budget of the Universe in ΛCDM, which is fortunate since it can be modelled relatively simply as a collisionless fluid that only

interacts through gravity. As a result, large volumes (of order ~ 1 Gpc^3) can be simulated at high resolution reasonably cheaply with current computational capabilities.

The basic approach is to divide the mass within some comoving volume, using periodic boundaries, up into particles of equal mass that interact through Newtonian gravity (GR corrections are negligible and typically ignored). The size evolution of the box is then found using Friedmann's equation (Friedman, 1922). Calculating the force between all N particles is an $\mathcal{O}(N^2)$ operation, so to reduce this cost simplifying approaches are used. The tree code uses a hierarchical octree algorithm, that essentially divides the volume into a hierarchy of cubic cells, allowing the calculation of long range forces to groups of particles to be done simultaneously, whilst still calculating individual interactions with nearby particles (Barnes & Hut, 1986). An alternative is the particle-mesh (PM) approach, which discretises space and calculates the potential on this grid using a Fourier transform of the sampled density field (Hockney & Eastwood, 1988). Both scale as $\mathcal{O}(N \log N)$, though PM has the advantage of being able to account for periodic volumes by default, whilst tree codes must use additional techniques such as Ewald summation (Hernquist et al., 1991).

All of the simulations used in this thesis use a hybrid tree-PM approach (e.g. GADGET-2, Springel et al., 2005), which combines the tree code for short range forces, and the PM approach for long range and periodic forces.

It is not feasible to output the full particle information at every time step.[1] In the Millennium simulation 64 snapshots were outputted from $z = 127$ to $z = 0$ spaced approximately linearly with the expansion factor (Springel et al., 2005).

2.3.2 Structure finding

Self-gravitating dark matter in an expanding universe proceeds to form a hierarchy of structures. The primary unit of structure is the virialised dark matter halo. Sheets, filaments and voids also make up a significant fraction of the total cosmic volume,[2] however it is in the collapsed, virialised halos that the majority of the mass resides; these

[1] A single snapshot from the fiducial EAGLE simulation (full hydro, box length 100 Mpc) occupies approximately 0.5 terabytes.

[2] Algorithmic approaches for identifying these structures in simulations have been developed, such as the Delaunay Tesselation (see Sousbie, 2011).

are the primary unit of structure in N-body simulations. I use a redshift independent definition of halo mass, given as the mass enclosed within a sphere with average density equal to 200 times the critical density of the universe.

To identify halos in dark matter simulations a number of different structure finders have been developed. The most conceptually simple and computationally cheap approach is Friends-Of-Friends (FOF, Davis et al., 1985; Efstathiou et al., 1985), which has been used to find structure in both observed galaxy distributions as well as numerical simulations (*e.g.* Farrens et al., 2011). A FOF halo is defined as a group of particles where the maximum separation of a given particle with all others is less than the linking length, l, where the value of l determines the size of structure found; hierarchies of structure can be defined using multiple linking lengths.

One drawback of the FOF approach is that it can link together disparate structures connected by a single link at the edge, which can be a drawback for finding unique collapsed structures. To overcome this other post-processing methods have been developed in order to identify a hierarchy of structures. The simulations used in this thesis all use the SUBFIND algorithm (Springel et al., 2001; Dolag et al., 2009). SUBFIND starts by using the FOF outputs; in hydro simulations this is done on the dark matter only, and gas and star particles are then associated with the nearest dark matter particle. It then calculates the local density at each particle position using an adaptive kernel estimation, assuming some number of neighbours for smoothing. In simulations with multiple particle species (*e.g.* hydro) this step is done on each species individually, and the total density is then the sum of the densities for each species. SUBFIND then defines peaks in the density as individual structures, which extend out to where the density reaches a saddle point with a neighbouring peak. Finally, a gravitational unbinding procedure is carried out to leave only the self-bound part remaining.

Halo catalogues at multiple different output times can be linked together to build *merger trees*, which describe the merger history of halos in simulations. These are typically built by finding common particles between halos in subsequent snapshots, and then linking these together to form a tree structure, where a halo can have multiple 'parent' halos in a previous snapshot. In some pathological cases halos can pass through each other without merging (*e.g.* the bullet cluster, Markevitch et al., 2004), which can lead to mis-classified

merger events. One means of avoiding this is to use 6D structure finders, that utilise not only the spatial but the velocity information of particles to identify bound structures in phase space (e.g. ROCKSTAR & VELOCIraptor, Behroozi et al., 2013a; Elahi et al., 2019). The choice of halo finder has only a small effect on individual halo properties and the cumulative halo and stellar mass function at $z = 0$ (Knebe et al., 2011, 2013), though there have been suggestions that high redshift structures, due to their clumpiness, are more dependent on the halo finder chosen (Klypin et al., 2011).

Each simulation used in this thesis uses a different code for constructing merger trees, which can lead to differences in the final trees (Srisawat et al., 2013). The fiducial EAGLE simulation uses the D-Trees code (Jiang et al., 2014; Qu et al., 2016), whereas the more recent Spiderweb algorithm has been applied to the C-Eagle simulations (Bahé et al., 2019). The Illustris simulation uses the Sublink algorithm (Rodriguez-Gomez et al., 2015), which has some modifications to the merger tree code described (but not explicitly named) in Springel et al. (2005), which has also been applied to the Millennium simulation, and used to construct the merger tree for the L-Galaxies SAM.

2.3.3 Semi-analytic models

SAMs are built on the outputs of dark matter only simulations, and use the merger trees coupled with differential equations describing the evolution of baryons in host halos (Baugh, 2006). A number of SAMs have been developed to model galaxy evolution at both low- and high-redshift (White & Frenk, 1991; Cole et al., 2000; Somerville et al., 2008; Gonzalez-Perez et al., 2014; Croton et al., 2016; Poole et al., 2016; Yung et al., 2018; Lagos et al., 2018). In this thesis I use L-galaxies, or the Munich SAM, described below.

2.3.3.1 The L-Galaxies Model

The latest version of L-Galaxies is an update to that presented in Guo et al. (2011) that uses the Planck 2013 cosmological parameters (Planck Collaboration et al., 2014), and better predicts the abundance of low-mass galaxies at $z \geq 1$ (Henriques et al., 2015). Using the abundance and passive fractions of galaxies at $z \leqslant 3$ the SAM model parameters are constrained using a Markov Chain Monte Carlo (MCMC) approach (Henriques et al., 2009; Lu et al., 2011; Henriques et al., 2009), which reproduces key observables during

this epoch such as the galaxy stellar mass function and optical luminosity functions. Despite being tuned to low redshift observables, the model also shows good agreement with high redshift galaxy properties, such as the stellar mass and luminosity function, out to $z = 7$ (Clay et al., 2015). A full description of the model is provided in the appendix to Henriques et al. (2015).

The growth of supermassive black holes is modelled in L-Galaxies through two mechanisms (Croton et al., 2006; Henriques et al., 2015). The first, labelled *quasar mode* growth, is triggered by a galaxy merger. The black holes merge instantaneously, and are then fed cold gas driven toward the nuclear region of the galaxy by turbulent motions induced by the merger. The second mechanism, labelled *radio mode* growth, is fed by hot gas from the halo, and leads to the formation of hot bubbles and jets. The quasar mode is the most effective mechanism by which black holes grow in the model, though the accretion is not explicitly associated with any feedback, except through supernovae feedback associated with the post-merger starburst in the case of a gas rich merger. In contrast, radio mode feedback leads to negligible black hole growth but produces efficient feedback that prevents the infall of cold gas in the largest halos.

2.3.4 Hydrodynamic Simulations

Hydrodynamic simulations differ from N-body simulations in that they explicitly and self-consistently model the evolution of gas and stars as well as dark matter. This allows the investigation of the spatial and kinematic distribution of these components, as well as their hydrodynamic and thermal interaction. A limitation to this approach is the resolution that can be simulated computationally efficiently, with 'subgrid' models required below the resolution limit. These subgrid models are analogous to the Semi-Analytic approach, but applied to much smaller scales, and require similar tuning of free parameters, though the computational cost of this tuning procedure is significantly greater. A number of hydro sims have been developed, including the MUFASA & SIMBA simulations (Davé et al., 2016, 2019), as well as models dedicated to studying the high redshift universe such as the Bluetides simulation, which simulates a large volume (400 Mpc box length) run down to $z = 8$ only (Feng et al., 2015a,b). Below I describe the hydrodynamic solvers and subgrid models used in the two hydrodynamic simulations used in this thesis: Eagle (Schaye et al., 2014; Crain et al., 2015) and Illustris (Vogelsberger et al., 2014; Genel et al.,

2014).

2.3.4.1 Hydrodynamic solvers

Smoothed Particle Hydrodynamics (SPH) is the most popular Lagrangian method. It models the mass distribution as discrete particles, with a kernel weighting determining the distribution of their physical properties (Springel, 2010a). The value of some field or quantity F at some arbitrary position is then given by the sum over the contribution from all neighbouring particles within some smoothing length, h,

$$X_i = \sum_j \frac{m_j}{\rho_j} F_j \, W(|\mathbf{r_i} - \mathbf{r_j}|, h) \ ,$$

where m is the particle mass, ρ its density, and W is some spherical kernel. One of the advantages of SPH is that it explicitly tracks the mass, which can be particularly useful in galaxy evolution codes for tracking the transfer of matter in outflows and inflows. However, SPH has traditionally suffered from artificial pressure boundaries, though modern codes are capable of mitigating this problem (for more details see Somerville & Davé, 2015).

Another approach is to take an Eulerian formalism and model the mass distribution on discrete cells, computing the advection of properties and forces across the boundaries of cells. Where the mass is high these can be refined to arbitrarily high resolution, an approach known as Adaptive Mesh Refinement (AMR), which leads to good adaptivity.

Hybrid codes, that combine the advantages of Lagrangian and Eulerian approaches, have recently become popular in cosmological hydrodynamic applications (Somerville & Davé, 2015). AREPO is an example of a Largrangian-Eulerian method, that uses a Voronoi Tesselation to subdivide the volume around particles into space-filling polyhedra (Springel, 2010b). Properties are then advected along the face of each cell boundary.

2.3.4.2 The Eagle simulations

Eagle is a recent hydrodynamic simulation from the Virgo consortium, that uses the SPH approach; the numerical methods are collectively called the 'Anarchy' suite, described in more detail in Schaller et al. (2015). The subgrid recipes are based on the OWLS suite used in GIMIC (Crain et al., 2009). Full details are provided in Schaye et al. (2014) and

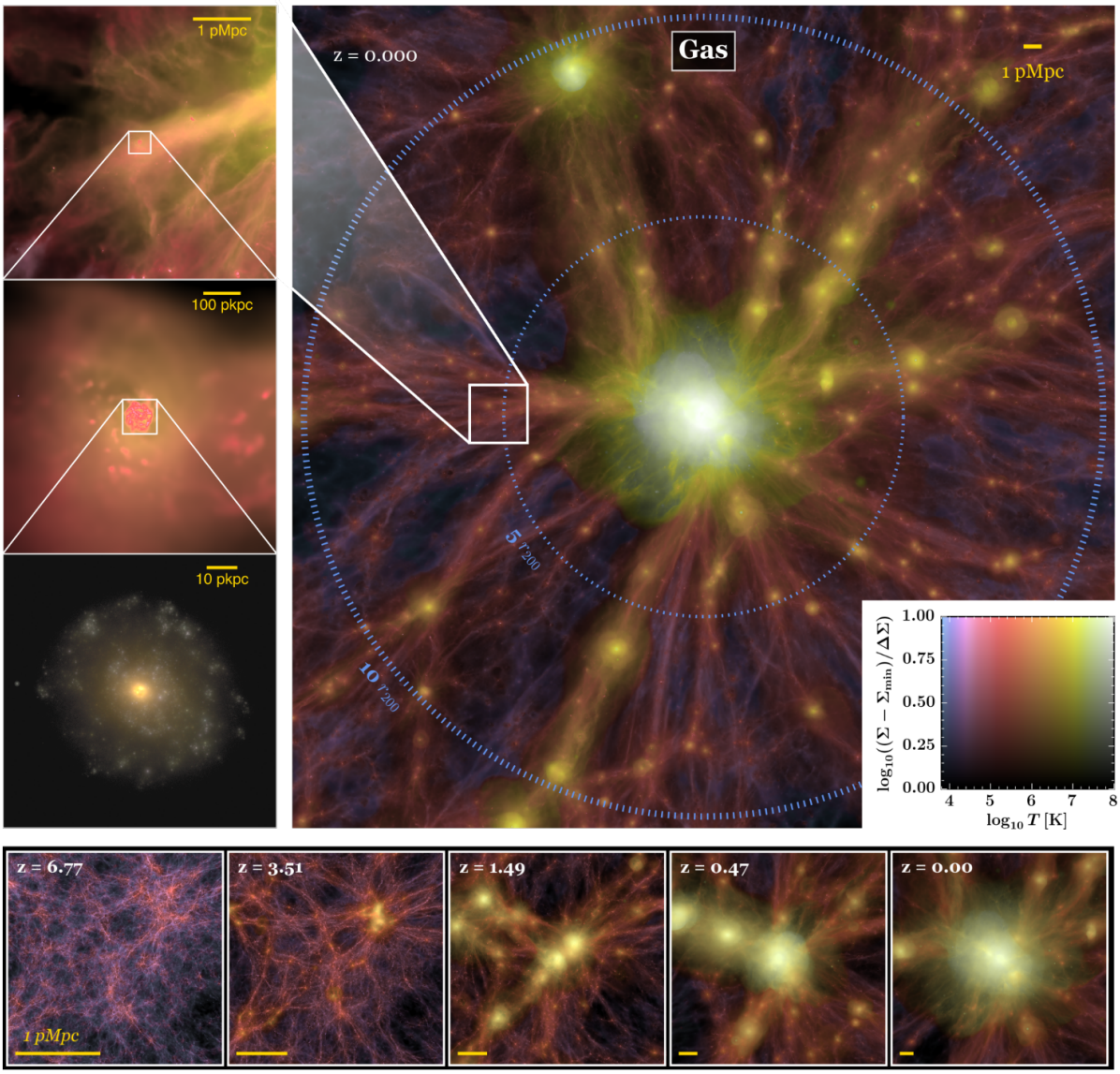

Figure 2.4: *Top panels:* the gas distribution at redshift $z = 0$ centred on a massive cluster ($M_{200} \,/\, M_{\odot} = 10^{15.38}$) from the C-EAGLE simulations, in a $60 \times 60 \times 15$ physical Mpc slice. Gas surface density is represented by brightness, and temperature by the colour (see HSV map in the bottom-right corner). *Top-left panels:* zoom in towards an individual galaxy; a synthetic *gri* image of the stellar content of the galaxy is shown in the bottom panel. *Bottom panels:* redshift evolution of the gas distribution. The diffuse web of filaments connecting dense nodes in the high-redshift ($z \geqslant 1.5$) protocluster environment is clearly visible. Reproduced from Bahé et al. (2017).

Crain et al. (2015); below I summarise the main components:

- Element-by-element radiative cooling recipes and photoheating following Wiersma et al. (2009a).
- A spatially-uniform ionizing UV background (UVB) turned on at $z = 11.5$ that then follows the time-dependent model of Haardt & Madau (2012).
- The pressure dependent star formation rate from Schaye & Dalla Vecchia (2008) with a metallicity dependent threshold from Schaye (2004).
- Stellar mass loss based on Wiersma et al. (2009b), whereby metals (and the associated transfer of momentum and energy) are distributed to neighbouring gas particles within the SPH kernel, the fraction depending on their relative distance (assuming star particles represent simple stellar populations with a Chabrier IMF; Chabrier, 2003).
- Stochastic thermal feedback from stars, following Dalla Vecchia & Schaye (2012), with metallicity and density dependent thermal losses.
- Black hole (BH) seeding in FOF halos with mass $> 10^{10}\,\mathrm{M}_\odot\,/\,h$ that do not already contain a BH (as in Springel et al., 2005), replacing the highest density gas particle.
- BH accretion, dependent on the BH mass, its relative velocity, and the local density, temperature and angular momentum of the surrounding gas.
- Thermal, stochastic AGN feedback proportional to the BH accretion rate (see below for details).
- Single-mode thermal AGN feedback with fixed efficiency (analogous to 'quasar mode' feedback in L-Galaxies) as in Booth & Schaye (2009).

These subgrid models were tuned to the following key distribution functions at $z = 0$: the GSMF, the SHMR, and the black hole-stellar mass relation, as well as galaxy sizes.

In detail, the BH accretion is given by the minimum of the Eddington and Bondi-Hoyle rates (Bondi & Hoyle, 1944), times some efficiency factor,

$$\dot{m}_{\mathrm{BH}} = (1 - \epsilon_r)\,\min(\dot{m}_{\mathrm{Edd}}, \dot{m}_{\mathrm{Bondi}} \times A) \tag{2.8}$$

where

$$\dot{m}_{\rm Edd} = \frac{4\pi G m_{\rm BH} m_p}{\epsilon_r \sigma_t c} \tag{2.9}$$

$$\dot{m}_{\rm Bondi} = \frac{4\pi G^2 m_{\rm BH}^2 \rho}{(c_s^2 + v^2)^{3/2}} \tag{2.10}$$

$$A = \min(C_{\rm visc}^{-1}(c_s/V_\phi)^3,\, 1) \ , \tag{2.11}$$

and m_{BH} is the black hole mass, m_p is the proton mass, $\epsilon_r = 0.1$ is the accretion disk radiative efficiency, σ_T is the Thomson cross section, ρ is the surrounding gas density, c_s is the sound speed, v is the relative velocity of the BH to the surrounding gas, and the other parameters have their usual meanings. V_ϕ is the rotation velocity of the gas around the black hole (Rosas-Guevara et al., 2015), and $C_{\rm visc}$ parametrises the viscosity of the accretion disc in the subgrid regime.

The fiducial Eagle simulation was a (100 Mpc)^3 periodic volume using the Reference (Ref) parameter set. This volume was sufficiently large to contain four cluster-mass halos at $z = 0$, allowing the investigation of galaxy evolution in a wide variety of environments. This was combined with a number of smaller, periodic volumes using different parameters. The AGNdT9 model, which is more sensitive to the gas viscosity in the surroundings of the SMBH, and injects energy less frequently in more energetic bursts, was run in a (50 Mpc)^3 volume, and showed better agreement with the observed gas mass-fraction and X-ray luminosity temperature of group mass objects ($M_{500}\,/\,M_\odot \sim 10^{13.5}$). The parameter differences between Ref and AGNdT9 are shown in Table 2.1.

Prefix	$C_{\rm visc}$	$\Delta\, T_{\rm AGN}\,(K)$
Ref	2π	$10^{8.5}$
AGNdT9	$2\pi \times 10^2$	10^9

Table 2.1: Subgrid parameter differences between the two Eagle models used in this thesis, the fiducial Reference model (Ref) and AGNdT9. $C_{\rm visc}$ controls the sensitivits of the black hole accretion rate to the angular momentum of the surrounding gas (see equation 2.11). and $\Delta\, T_{\rm AGN}$ is the value of gas temperature increase during an episode of AGN feedback.

Periodic volume simulations necessarily simulate a mean-density patch of the universe. In order to capture extremes of the overdensity distribution large volumes must be simulated to capture the largest modes in the power spectrum, which are currently computationally

unfeasible with high resolution hydrodynamics. An alternative is to carry out 'zoom' simulations, which re-simulate a patch of a large, low resolution simulation at high resolution, typically with added physics such as hydrodynamics (Katz & White, 1993; Tormen et al., 1997). The surrounding region is then represented by low resolution dark matter only particles that interact only gravitationally, but preserve the large scale power. Figure 2.5 shows a cartoon demonstration of this approach. This was the approach used in the C-Eagle simulation suite, which re-simulated 30 clusters of a range of descendant halo masses with the full Eagle hydrodynamic model, at fiducial resolution with the AGNdT9 model parameters. The Clusters were selected from a $(3.2\ \mathrm{Gpc})^3$ dark matter only simulation (Barnes et al., 2017a), with particle resolution $m_{\mathrm{DM}} / M_{\odot} = 5.43 \times 10^{10}\, h^{-1}$. The high resolution region was re-simulated using the same gas and dark matter particle masses as in the fiducial Eagle AGNdT9 simulation to facilitate comparison. C-Eagle shows good agreement with central black hole and total stellar mass estimates, but some discrepancies in the gas and Brightest Cluster Galaxy (BCG) masses at $z = 0$; full details are provided in Barnes et al. (2017b); Bahé et al. (2017).

2.3.4.3 The Illustris simulations

The Illustris simulation uses the hybrid Arepo hydro scheme (Springel, 2010b), with a maximum resolution at $z = 0$ of 48 pc. The subgrid models are described in detail in Vogelsberger et al. (2013); below I summarise its main components:

- A spatially uniform time-dependent UV background from Faucher-Giguère et al. (2009).
- Metal line cooling using Cloudy tables (Ferland et al., 2013), analogous to Wiersma et al. (2009a) using identical elements (H, He, C, N, O, Ne, Mg, Si, S, Ca, Fe).
- An on-the-fly self-shielding prescription from the UVB based on Rahmati et al. (2013).
- Kinetic, stochastic stellar-feedback decoupled from the hydro scheme, as in Springel & Hernquist (2003); Oppenheimer & Davé (2008).
- A density dependent star formation rate from Springel & Hernquist (2003) with an upper temperature ceiling, assuming a Chabrier IMF (Chabrier, 2003).

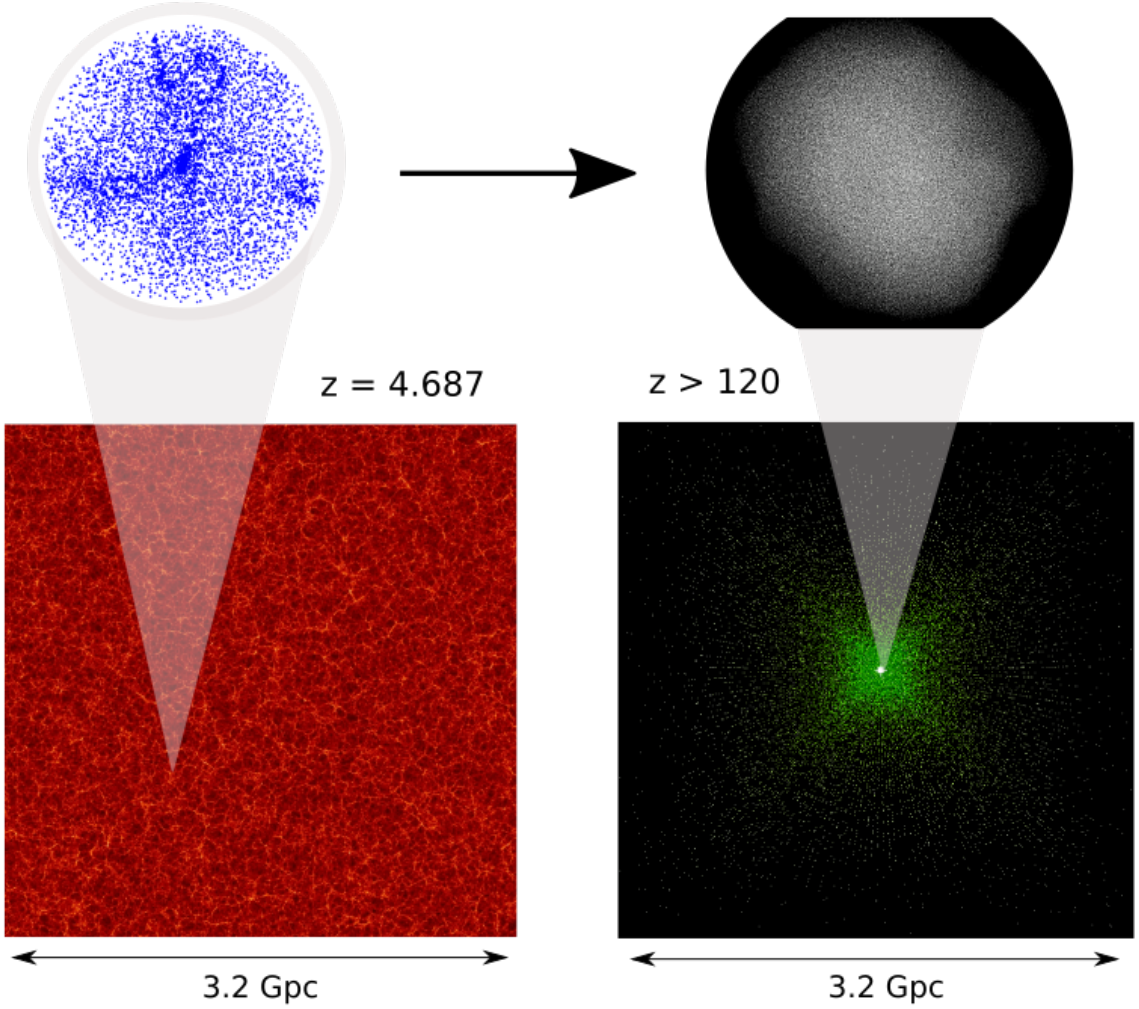

Figure 2.5: A cartoon showing the selection of an overdense region from the low resolution dark-matter only simulation, and its re-simulation at high resolution in a zoom simulation. In this example the selection is made at high redshift ($z = 4.687$), whereas in C-Eagle clusters are selected at $z = 0$. The selection is re-centred in the box, and a hierarchy of low resolution dark matter only particles form a 'glass' around the high resolution region.

- A stellar mass loss recipe similar to Wiersma et al. (2009b), where instead of neighbouring particles, neighbouring Voronoi cells are enriched over a tophat kernel.
- Black hole seeding in FOF halos with mass $> 5 \times 10^{10} \mathrm{M}_{\odot} \, / \, h$, replacing the highest density gas cell.
- Bondi-Hoyle-Lyttleton Eddington-limited SMBH accretion (see Edgar, 2004).
- Two-mode AGN feedback, consisting of a thermal ('quasar') and mechanical ('radio') mode from Sijacki et al. (2007) as well as an additional electromagnetic feedback component; the mode is determined by the accretion rate.

The Illustris subgrid model was tuned to two observed relations: the SHMR at $z = 0$, and the CSFRD. The fiducial Illustris simulation was run on a $(106.5 \text{ Mpc})^3$ periodic volume.

Illustris-TNG is a recent update to the Illustris model that implements, among other updates and improvements, the effects of magnetic fields (magneto-hydrodynamics) (Pillepich et al., 2017; Donnari et al., 2019).

2.4 Spectral Energy Distribution Modelling

Further post-processing is required to link the intrinsic properties of galaxies to observables in cosmological simulations. The main components of the SED modelling pipeline are the emission from stars, attenuation by dust, the attenuation and emission from nebular regions, and the contribution from AGN (Conroy, 2013). There are a number of approaches for modelling each component, of varying complexity. Here I will describe the basic principles, the dust-screen approach used in this thesis, as well as more advanced approaches for future investigation.

2.4.1 Population Synthesis

The emission from a single star can be written

$$f_{\lambda}(m, t, Z) \ ,$$

where m is the initial mass, Z its metallicity, t its age, and f_{λ} its flux at wavelength λ. The presence of a binary companion can also significantly affect the resulting emission,

discussed below.

Star elements in EAGLE and Illustris represent of the order of $10^6\,M_\odot$ of stars, much larger than average individual star clusters. It is therefore necessary to model their emission as if they represent Single Stellar Populations (SSP) with uniform age and metallicity, integrated over the IMF,

$$f_\lambda^{\rm SSP} = \int_{M_{\rm min}}^{M_{\rm max}} \phi_{\rm IMF}(M) f_\lambda(M,t,Z) {\rm d}M \quad ,$$

where ϕ is the assumed IMF, and $M_{\rm max}$ & $M_{\rm min}$ are the upper and lower IMF mass limits.

The details of $f_\lambda^{\rm SSP}$ are modelled using Stellar Population Synthesis (SPS) models. A number of SPS codes have been developed, including PEGASE (Fioc & Rocca-Volmerange, 1997, 1999, 2019), BC03 (Bruzual & Charlot, 2003), M05 (Maraston, 2005) and FSPS (Conroy et al., 2009; Conroy & Gunn, 2010).

Recently, the contribution from binary stellar populations was included in the BPASS models (Stanway et al., 2016; Eldridge & Stanway, 2016; Eldridge et al., 2017; Stanway & Eldridge, 2018). This has a significant impact on the ionising emission, with important repercussions for the Epoch of Reionisation (EoR) (Wilkins et al., 2013b). A consequence of including binaries is a dramatic increase in the parameter space of the SPS model, in order to take account of the impact of, for example, their separation, mass ratio, multiplicity, *etc.*, and how these are distributed for a given SSP. The impact of certain phases of a stars evolution are also theoretically uncertain, such as the Thermally Pulsating-Asymptotic Giant Branch (TP-AGB) phase, which, when omitted or modelled incorrectly in SPS models can result in systematic differences in the resulting emission, and hence the derived physical parameters (Conroy et al., 2009).

Figure 2.6 shows the predicted spectral energy distribution from the UV to the NIR for five different SPS models at a given age and metallicity. Figure 2.7 shows the effect of varying the age and metallicity for the FSPS model.

It is computationally unfeasible to calculate the exact SPS emission from each SSP, so grids of age and metallicity are pre-calculated, and the emission from a given SSP is estimated through 2D interpolation. The total stellar emission from a galaxy is then

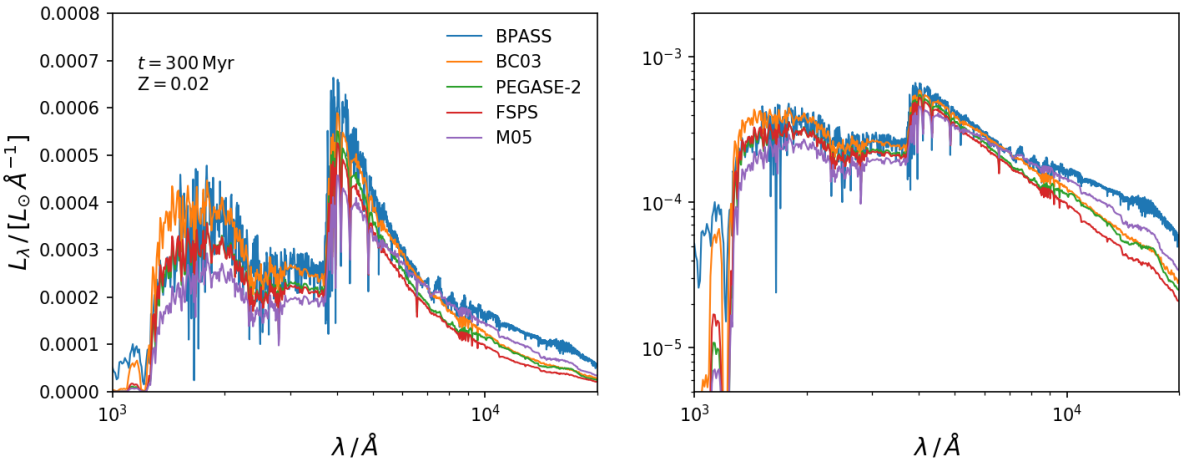

Figure 2.6: Spectral energy distribution for a simple stellar population with age 300 Myr and metallicity Z = 0.02, for five different SPS models, in linear- (left) and log-space (right).

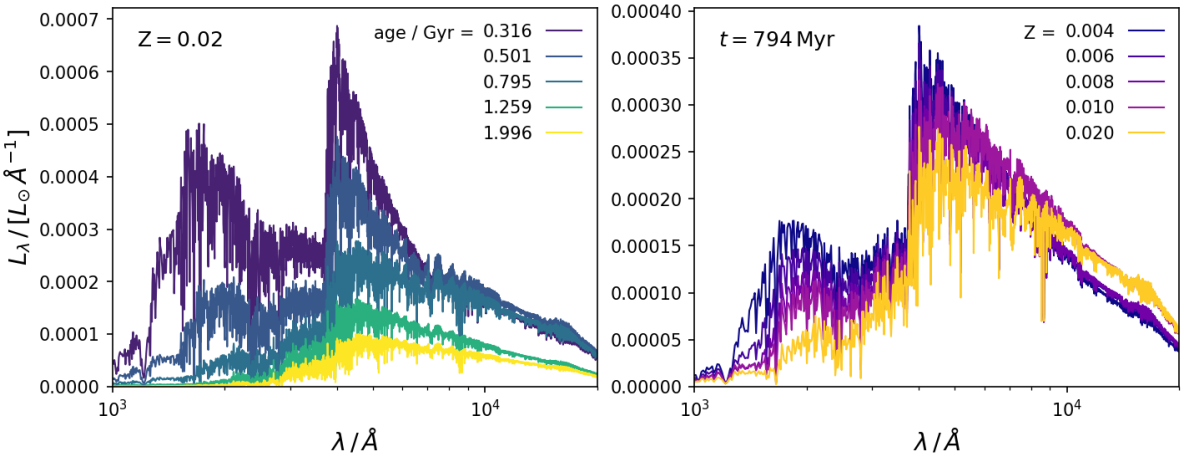

Figure 2.7: Spectral energy distribution from FSPS with varying parameters. *Left:* Varying age, with a fixed metallicity of Z = 0.02. *Right:* Varying metallicity, with a fixed age of 794 Myr.

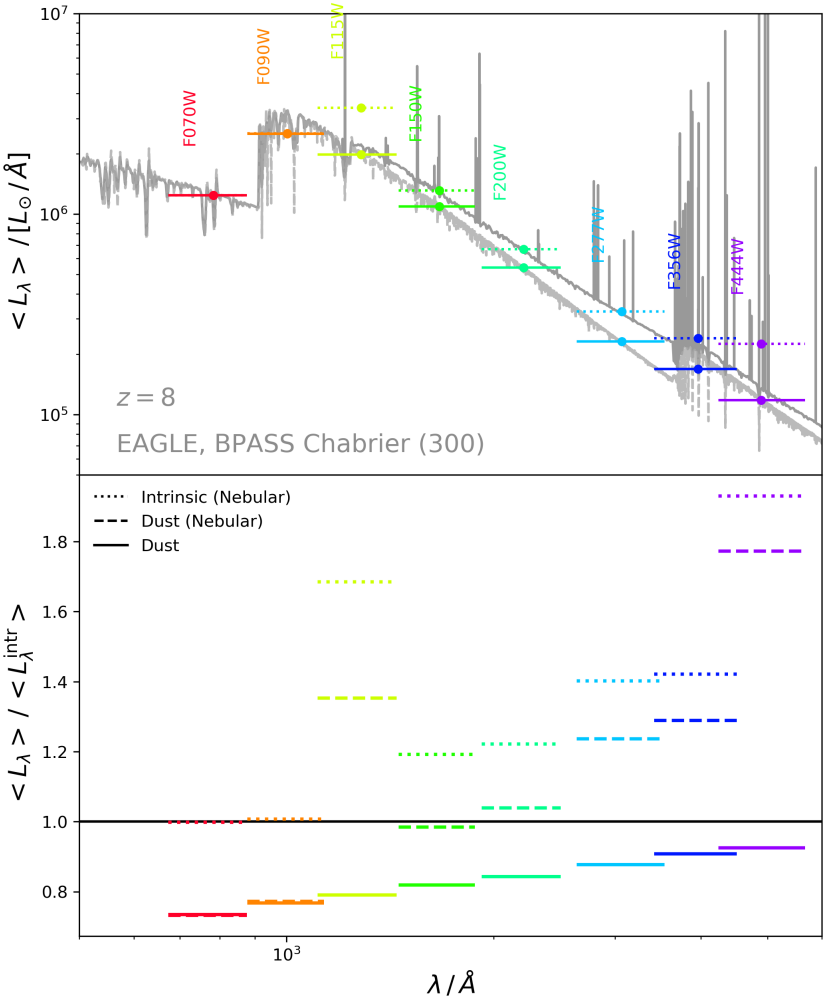

Figure 2.8: Mean SED from the fiducial EAGLE simulation at $z = 8$. *Top:* the light grey shows the intrinsic distribution, the darker grey includes the nebular component. The response in the JWST NIRCam filters is shown by the coloured lines (solid for intrinsic, dotted including nebular). *Bottom:* the ratio of flux in the JWST NIRCam filters to the intrinsic flux for different modelling assumptions.

obtained by combining the emission from each stellar element in the simulation to give the composite stellar spectrum, which given the grids is a simple matrix multiplication and addition operation. Figure 2.8 shows the average SED from the fiducial EAGLE simulation at $z = 8$.

2.4.2 Dust attenuation

Dust grains are tiny solid particles, approximately 0.1 μm across, consisting of a variety of elements present in the ISM. Despite making up only 1% of the ISM mass, dust is responsible for reprocessing around 30% of all photons in the universe (Bernstein et al., 2002), scattering and absorbing it at wavelengths below its size, which increases the dust temperature, then re-emitting that radiation at mid- and far-infrared wavelengths. Dust is produced and destroyed through a complex network of physical processes in the ISM, leading to a non-trivial dependence on the galaxies SFH and ISM evolution (Vijayan et al., 2019).

A simple model for the dust attenuation is a 'slab' in front of the stellar source populations that attenuates as

$$f_{\rm obs}(\lambda) = f_{\rm int}(\lambda)\, 10^{0.4\, A_\lambda} \quad ,$$

where $f_{\rm int}$ & $f_{\rm obs}$ are the intrinsic and observed flux, respectively, and A_λ is a linear scaling parameter, given by

$$A_\lambda = k(\lambda)\, E(B\, -\, V) = \frac{k(\lambda) A_V}{R_V} \quad ,$$

where $k(\lambda)$ is the reddening curve, $E(B - V)$ traditionally refers to the extinction between the B and V photometric bands, and $R_V = A_V \,/\, E(B\, -\, V)$. In practice, A_V and R_V are simply parameters of the dust extinction model, describing the normalisation and shape, respectively (Barbary, 2016a). Some popular parametrisations of $k(\lambda)$ include those by O'Donnell (1994); Fitzpatrick (1999); Calzetti et al. (2000), shown in Figure 2.9.

In cosmological simulations, the magnitude of the attenuation can be linked to the physical parameters of the galaxy. The mass of metals in the ISM is closely linked to the mass of dust; this is often used as a proxy, by taking the total metallicity of all gas elements multiplied by their mass, times some constant (Trayford et al., 2015; Narayanan et al., 2017). Many screen dust models include a second component that additionally attenuates

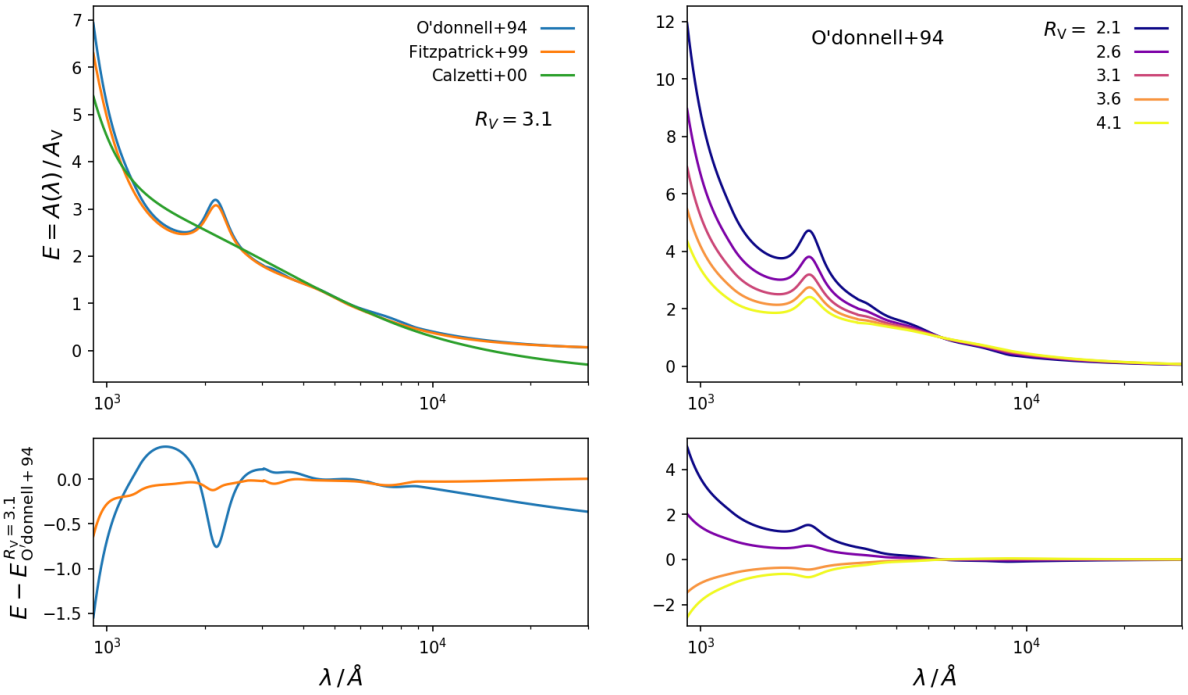

Figure 2.9: *Left:* dust extinction curve parametrisations from O'Donnell (1994); Fitzpatrick (1999); Calzetti et al. (2000). *Right:* the affect of changing R_V on the extinction curve (using O'Donnell, 1994).

young stellar populations < 10 Myr in age (Trayford et al., 2015). This is physically motivated by the excess of dust in nebular systems. Similar screen models have been implemented in a number of SAMs (Vijayan et al., 2019).

A more advanced approach is to calculate the line-of-sight (LOS) attenuation to each star particle. In SPH simulations this is typically done by integrating the kernel for each gas particle within some impact parameter (*e.g.* LOSER, Davé et al., 2017). The gas density can then be linked to the metallicity, and hence the dust attenuation, as above (Wilkins et al., 2016). This more physical approach takes into account the spatial distribution of the dust with respective to the stars.

Further details on the dust screen model used in this thesis are described in Chapter 5.

2.4.3 Nebular Contribution

Massive stars in young stellar populations emit abundant Lyman continuum photons, which ionise their surrounding gas to produce nebular line and continuum emission, as well as nebular attenuation. The emission and attenuation in these regions is typically modelled using spectral synthesis codes utilising photo-dissociation models, such as

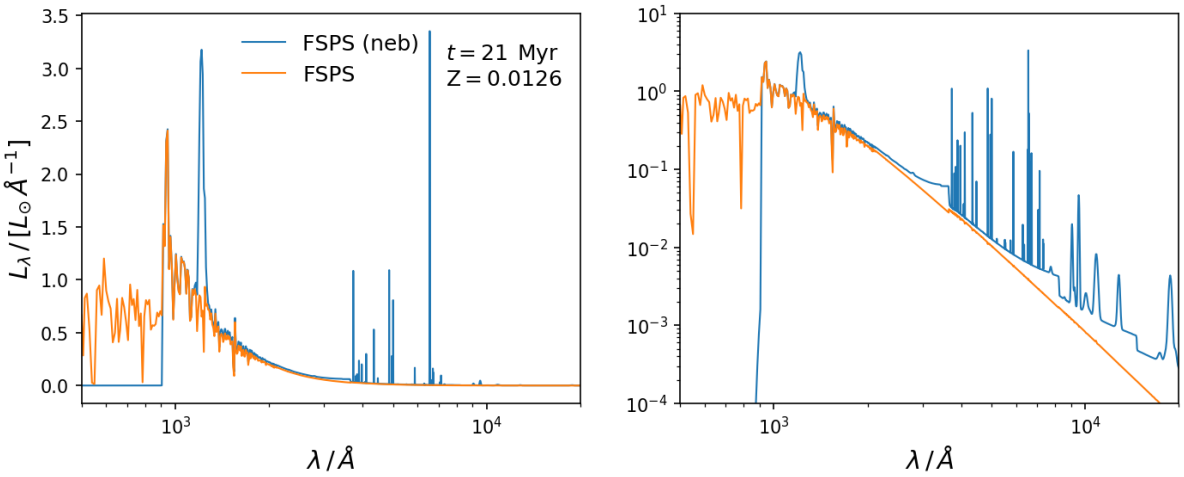

Figure 2.10: Spectral energy distribution for a simple stellar population with age 21 Myr and metallicity Z = 0.0126 from FSPS (Conroy et al., 2009; Conroy & Gunn, 2010), both with and without nebular attenuation according to the prescription in Byler et al. (2017).

CLOUDY (Ferland et al., 2017).[3] FSPS (Conroy et al., 2009; Conroy & Gunn, 2010) now includes a self-consistent nebular component (Byler et al., 2017) that links the ionising radiation from the SSP to the ionisation rate of the surrounding cloud using the dimensionless ionisation parameter U,

$$U = \frac{Q_{\rm H}}{4\pi R^2 \, n_{\rm H} \, c} \, , \tag{2.12}$$

where R is the radius of the ionized region, $n_{\rm H}$ is the hydrogen number density, and $Q_{\rm H}$ is the ionising emissivity,

$$Q_{\rm H} = \frac{1}{hc} \int_0^{\lambda_0} \lambda \, f_\lambda \, d\lambda \tag{2.13}$$

where $\lambda_0 = 921$Å is the Lyman limit. This parametrisation folds in both the geometry and hardness of the ionising spectrum, reducing the dimensionality, and hence the size of the grid of models needed.

2.4.4 Radiative Transfer Approaches

There are also more complex approaches for estimating the effect of dust that not only take into account the spatial distribution of dust in the galaxy and its affect on attenuation,

[3] https://www.nublado.org

but also scattering and re-emission. These radiative transfer approaches explicitly model the release of photons from source stellar populations and their interaction with gas and dust in the ISM in a full 3D geometry. SKIRT is one example of a radiative transfer code, and has been applied to the EAGLE simulation to self-consistently predict the Far-Infrared (FIR) emission of galaxies (Trayford et al., 2017; Camps et al., 2016).

However, these approaches are still applied in the post-processing phase to individual snapshots. Cosmological models typically do not model radiative processes self-consistently, instead using subgrid models to describe the effects of *e.g.* photoionisation of gas by young stars. A more self-consistent approach applies radiative transfer during running of the simulation. These radiative transfer cosmological models are computationally expensive, prohibitively so for large cosmological volumes at sufficient resolution to resolve the properties of galaxies self-consistently, but have been successfully applied to zoom simulations of the EoR (Iliev et al., 2006, 2009).

2.5 Galaxy Protoclusters

Galaxy clusters are rare collections of hundreds, sometimes thousands of galaxies embedded within a massive ($> 10^{14}\,\mathrm{M}_{\odot}$), virialised, dark-matter halo. They are observable through extended X-ray emission from a hot intracluster medium, but can also be identified from the red colours and elliptical morphologies of their constitituent galaxies compared to non-cluster (field) galaxies (Dressler, 1980; Vikhlinin et al., 2014). Protoclusters are the pre-collapse progenitors of clusters, commonly defined as the ensemble of objects that will collapse into the cluster by $z = 0$ (Overzier, 2016). They manifest as matter overdensities with respect to the field extended over large comoving volumes (up to 40 comoving Mpc across at $z \sim 3$; Muldrew et al., 2015), which are observable through surface overdensities of galaxies (Chiang et al., 2013; Lovell et al., 2018).

2.5.1 Identifying Protoclusters

Observational searches for protoclusters tend to adopt one of two approaches: 'blind' searches for surface overdensities of galaxies, and focused observations around biased tracers. The former typically work by identification of surface overdensities in wide field photometric surveys of Lyman break galaxies (LBGs) and narrow band imaging of

emission line galaxies (Shimasaku et al., 2003; Adams et al., 2011; Spitler et al., 2012; Chiang et al., 2014; Shimakawa et al., 2017a), which are often followed up and confirmed spectroscopically (Toshikawa et al., 2012; Diener et al., 2015; Toshikawa et al., 2016). The VIMOS Ultra Deep Survey, the largest purely spectroscopic search, recently announced the discovery of a massive candidate at $z \sim 4.57$ (VUDS, Fèvre et al., 2015; Lemaux et al., 2017).

The second method takes advantage of objects thought to represent biased tracers of the underlying matter distribution, such as dusty star forming galaxies (Capak et al., 2011; Casey et al., 2014), Ly-α emitting blobs or extended Ly-α absorbers (Hennawi et al., 2015; Cai et al., 2016), High-redshift Radio Galaxies (HzRGs) and quasars. Using biased tracers to search for protoclusters is cheaper than performing wide, deep surveys. However, the uncertainty in their correlation could arguably make them unreliable: they may not probe a significant fraction of protoclusters (Orsi et al., 2016), or produce an unrepresentative sample of the population.

Given a galaxy overdensity measured with one of the above approaches, we wish to know the probability that it represents a protocluster, and an estimate of its descendant cluster mass, a useful property on which many other protocluster properties (size, maturity) depend. They can be estimated analytically (Steidel et al., 1998, e.g.), or from cosmological simulations (Suwa et al., 2006): protocluster probability is typically estimated by taking the ratio of regions with a given overdensity that evolve into protoclusters to those that do not (Chiang et al., 2013, 2014), and estimates of descendant mass have been inferred empirically from the typical descendant mass of a protocluster with similar overdensity (Orsi et al., 2016). Approaches such as these have been used in the construction of some of the first protocluster catalogues (Franck & McGaugh, 2016a,b; Higuchi et al., 2018).

Measures of overdensity are typically carried out with apertures or nearest neighbour approaches, the former showing greater correspondence with the actual 3D overdensity (Shattow et al., 2013), though orientation, aperture size and redshift uncertainty can have a significant effect on the quantitative overdensity value (Chiang et al., 2013; Monaco et al., 2005), which can in turn affect probability and mass estimates. In particular, redshift uncertainty acts to effectively elongate the measurement aperture, which lowers the measured overdensity by including more field volume. It also complicates the definition of

a protocluster in simulations - when does a randomly selected irregular aperture represent a protocluster or not? Prior to virialisation, protoclusters are an integral part of the high redshift cosmic web, tracing the nodes and filaments of the large scale structure (Overzier, 2016; Shimakawa et al., 2017a), which also complicates their identification and discrimination from the field, particularly so when using elongated apertures due to the risk of alignment.

2.5.1.1 Protoclusters traced by Active-Galactic Nuclei

A significant number of protoclusters have been found targeting HzRGs (Fèvre et al., 1996; Miley et al., 2006; Venemans et al., 2007; Galametz et al., 2010; Hatch et al., 2011a; Koyama et al., 2012; Wylezalek et al., 2013; Shimakawa et al., 2014; Cooke et al., 2014). Both Ramos Almeida et al. (2013) and Hatch et al. (2014) propose that the large-scale overdense environment may be causally connected to the presence of a radio-loud AGN, which may not necessarily reside at the peak of the overdensity. Searches surrounding quasars, on the other hand, have turned up a less conclusive picture; whilst many luminous quasars are clearly located in overdensities (Husband et al., 2013; Adams et al., 2015; Hennawi et al., 2015; Morselli et al., 2014; Mazzucchelli et al., 2017; Miller et al., 2016), many reside in average overdensity environments (Willott et al., 2005; Uchiyama et al., 2017).

2.5.2 Star Formation in Protocluster Environments

The SFS shows a strong dependence on environment at low redshift, with a clear red-sequence observed in local clusters, however Peng et al. (2010) find a more moderate overdensity dependence outside collapsed clusters. There are few comprehensive observational studies of protoclusters due to their rarity, large angular sizes and the difficulty of identifying their constituent galaxies from uncertain redshift estimates. However, a small number of protoclusters between $1.5 < z < 2.5$ have been studied in detail, with spectroscopic redshifts of star forming galaxies allowing accurate determination of their protocluster membership. These observations suggest a very small environmental dependence on the SFS (Koyama et al., 2013; Cooke et al., 2014; Duivenvoorden et al., 2016). For passive galaxies, determining the protocluster membership is more difficult, as they are typically only seen in photometry, which incurs large redshift uncertainties.

Where stellar mass and SFR estimates of the constituent galaxies have been derived, a similar trend to that at low redshift is emerging (Koyama et al., 2012, 2013; Shimakawa et al., 2018, 2017a); protoclusters contain a higher *density* of star forming galaxies, but the star formation rate for a galaxy at a given stellar mass is very similar to that seen in the field (Shimakawa et al., 2017a, 2018; Smith et al., 2019). However, the passive fraction of galaxies does exhibit a strong environmental dependence up to $z \sim 2.5$ (Lee-Brown et al., 2017; Cooke et al., 2016; Newman et al., 2014).

2.5.3 Numerical Studies of Protoclusters

Simulations allow us to explicitly follow the evolution of overdense environments, and link the properties of their descendant clusters to their high redshift progenitors. Due to the rarity of clusters, large volumes are needed to obtain a large sample, which necessitates either semi-analytic approaches or 'zoom' simulations of individual objects. Chiang et al. (2013) used a Semi-Analytic Model (SAM) to study the spatial distribution of protoclusters, and derived relationships between high redshift overdensities and the descendant cluster mass, a useful protocluster diagnostic. Contini et al. (2016) studied the properties of protoclusters in resimulations using a SAM, and found good agreement with observations of the star formation rate as a function of radius from the most massive galaxy. In Chapter 3 I extend the analysis of Chiang et al. (2013), finding an optimal aperture within which to measure galaxy overdensities in order to best identify and characterise them.

Simulations have found that the majority of the $z = 0$ cluster stellar mass is formed at $z > 3$, particularly the stellar content of the central brightest cluster galaxy (BCG), and then assembled at lower redshifts (Lucia & Blaizot, 2007; Ragone-Figueroa et al., 2018). Despite their rarity, protoclusters occupy an increasing fraction of the cosmic volume with redshift (up to 5% by $z = 7$) and contribute significantly to both the cosmic star formation rate and stellar mass density at high redshift (15% and 40% at $z = 2$ and $z = 7$, respectively; Chiang et al., 2017; Muldrew et al., 2018). It is therefore crucial to model these overdense environments at high redshift correctly in order to predict both high redshift observables and the low redshift properties of their collapsed descendants.

In Chapter 4 I use the high-redshift outputs of the Cluster-Eagle (C-Eagle) simulations (Barnes et al., 2017b; Bahé et al., 2017), 30 zoom simulations of galaxy clusters using the

EAGLE model, to study protocluster environments across a range of descendant cluster masses. I use the SPIDERWEB merger trees (Bahé et al., 2019) to identify protocluster galaxies, and treat the large sample of galaxies outside protoclusters as a comparison field region. I also utilise two periodic box simulations (Schaye et al., 2014; Crain et al., 2015) as additional comparison field regions, extracting any protoclusters contained in them. Together these simulations allow the study of the dependence of the SFS on protocluster environment.

2.6 Machine Learning Methods

Machine Learning (ML) methods have become increasingly popular in Astronomy in the last 20 years (Ball & Brunner, 2010; Baron, 2019), and can be divided into two main approaches: supervised and unsupervised learning. In this thesis I use both supervised and unsupervised approaches, described in detail in Chapter 5.

Supervised ML approaches use existing data to train a machine to predict on unseen data; this can take the form of a regression or classification problem. Compared to traditional model fitting approaches, ML methods do not require a parametrised model defined up front, instead building a flexible model directly from the input data. The input data are known as *features*, and the output data you wish to predict a relationship with are known as *predictors*. Where the predictors are continuous this is a regression problem; where they are discrete it is a classification problem. The features and predictors are split into training and testing sets, typically 80-20%, respectively. The training set is used to train and tune any hyperparameters of the model; the test set is held out, not involved in any of the training, and used at the end of the analysis to evaluate the model.

Unsupervised learning algorithms learn from data without being given known relationships between variables, and are typically used for clustering analysis, dimensionality reduction, outlier detection and visualisation.

Machine learning methods can have large numbers of free parameters. For Artificial Neural Networks (ANNs) one can think of the weights and biases of each 'neuron' as an individual parameter, which means large ANNs have tens of thousands of free parameters to tune and choose. The 'hyperparameters' are then considered the higher-level parameters of the model, such as the number of layers, or the chosen activation function. These parameters

are often chosen algorithmically through brute-force approaches on the training data, however rather than using the test set for evaluation, hyperparameter optimisation is carried out on a subset of the training data called the *validation* set, typically 10% of the training data.

It is often necessary to *normalise* the features to a common scale. This is necessary where there are multiple features with different dynamic ranges, or for certain ML approaches such as neural networks where the activation function is insensitive to features outside some given range (typically mean zero and variance one, for regression problems).

For supervised approaches the model is evaluated using some *loss function*, which evaluates the distance between the prediction and the true predictors. Typical loss functions for regression problems include the Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE). In Chapter 5 I present an alternative loss function, Symmetric Mean Absolute Percentage Error, which has particular advantages over other loss functions in the evaluation of step-wise SFHs, where the SFR is strictly $\geqslant 0$.

3 Characterising and Identifying Galaxy Protoclusters

Christopher C. Lovell,[1] Peter A. Thomas,[1] and Stephen M. Wilkins[1 4]

Accepted in Monthly Notices of the Royal Astronomical Society, 2017 November 25. Received 2017 November 25; in original form 2017 May 19.

3.1 Introduction

In this chapter we present a study of the characteristics of galaxy protoclusters using the latest L-GALAXIES semi-analytic model (Lovell et al., 2018), including an improved procedure for generating descriptive statistics of protoclusters that models the shape of the measurement aperture, and a robust protocluster definition for generating probabilities. We also investigate the spatial characteristics of protoclusters in order to determine whether the simplifying assumption of spherical symmetry is accurate, and how best to discriminate protoclusters from the field.

We describe our definitions, selection criteria and the L-GALAXIES model in Section 3.2, the galaxy population in protoclusters as a whole (Section 3.3.1), then characterise protoclusters in terms of their shapes (Section 3.3.2) and sizes (Section 3.3.3). We investigate the relationship between protoclusters and AGN in Section 3.3.5, and finally in Section 3.3.4 outline a procedure for generating improved statistics on galaxy overdensities, and apply the procedure to candidates from the literature (Section 3.4). Where we state comoving lengths, as opposed to physical , we precede the units with a lower case c, *e.g.* cMpc represents comoving mega parsecs.

[41] Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton, BN1 9QH, UK

3.2 Models and Methods

3.2.1 Simulation

We use the Millennium dark matter N-body simulation (Springel et al., 2005), which evolves 2160^3 particles (with mass $1.43 \times 10^9\, M_\odot$) from $z = 127$ to $z = 0$, in a comoving box with side length 480.3 h^{-1} cMpc. The original simulation was run using WMAP1 cosmological parameters (Spergel et al., 2003), however in this paper we use the halo properties rescaled to the Planck1 cosmology using the method described in Angulo & White (2010).

L-GALAXIES, or the Munich SAM, is a Semi-Analytic Model of galaxy evolution, described in greater detail in Chapter 2.

The AGN model in L-GALAXIES is a relatively simple phenomenological representation of the physical processes that lead to observable quasar and radio activity. It does not, for example, provide spin information, necessary for a complete description of the radio jet power (Fanidakis et al., 2011). As such, it does not match quantitative observational constraints on the accretion rate and black hole mass at high redshift. However, in this study we are primarily interested in the number density and spatial distribution of AGN and their hosts with regards to protoclusters; since AGN activity in the model depends explicitly on host halo mass, and implicitly on environment, a simple accretion cut should allow us to evaluate their coincidence with overdensities at high-z. A detailed description of AGN number densities, host halo masses and selection criteria is described in Section 3.3.5.

3.2.2 Definitions

We define as a *cluster* any Friends-of-Friends (FoF) halo at $z = 0$ with $M_{200}/M_\odot > 10^{14}$. Using this definition we identify 3825 clusters. We treat everything within R_{200} of the halo centre as a cluster member, and anything outside a cluster is labelled part of the *field*.

Throughout the paper, we use the following definition of a protocluster: that it is the ensemble of all objects that eventually end up in a present day cluster. Specifically, a

protocluster member is any halo or galaxy whose descendant at $z = 0$ lies within R_{200} of a cluster. To identify the protoclusters at a given epoch we follow the merger tree rooted on each subhalo in the cluster at $z = 0$, including the central subhalo, back in time to identify all progenitor halos and their galaxies.

3.2.3 Galaxy selection

We apply four galaxy selection criteria, identical to those employed in Chiang et al. (2013), with an additional high star formation rate selection, at snapshots corresponding approximately to $z = [2, 3, 4, 5, 6, 7, 8, 9.5]$:

$$\mathrm{S_{MAS9}}: \quad \log_{10}(M_*/\,M_\odot) > 9 \tag{3.1}$$

$$\mathrm{S_{MAS10}}: \quad \log_{10}(M_*/\,M_\odot) > 10 \tag{3.2}$$

$$\mathrm{S_{SFR1}}: \quad \mathrm{SFR}/(M_\odot\,\mathrm{yr}^{-1}) > 1 \tag{3.3}$$

$$\mathrm{S_{SFR5}}: \quad \mathrm{SFR}/(M_\odot\,\mathrm{yr}^{-1}) > 5\,. \tag{3.4}$$

The star formation rate selections ($\mathrm{S_{SFR1}}$ and $\mathrm{S_{SFR5}}$) most closely resemble the selection of line emission galaxies using narrow-band filters (e.g. Cooke et al., 2014).

3.2.4 Overdensity

Measures of protocluster overdensity using fixed volume apertures lead to greater consistency with redshift and better correspondence with the true 3D overdensity as compared to nearest neighbour approaches (Muldrew et al., 2012; Shattow et al., 2013). We define overdensity as

$$\delta_g(\boldsymbol{x}, V, z) \equiv \frac{n_g(\boldsymbol{x}, V, z)}{\langle n_g(V, z)\rangle} - 1\,,$$

where $\delta_g(\boldsymbol{x}, V, z)$ is the overdensity within a volume V centred on position $\boldsymbol{x}$ at redshift z. The volume can be spherical, $V = \frac{4}{3}\,\pi\,R^3$, or cylindrical, $V = \pi\,R^2\,D$, where R is the radius on the plane of the sky and D is the depth in the line-of-sight direction; we make clear in the relevant sections which volume is being used. $n_g(\boldsymbol{x}, V, z)$ is the number of selected galaxies within the chosen volume centred on $\boldsymbol{x}$, and $\langle n_g(V, z)\rangle$ is the mean number of selected galaxies in a volume of this size averaged over the entire simulation.

Where we wish to compare measured overdensities as closely as possible to observations, we must take into account peculiar motions along the Line-of-Sight (LoS). High velocities along the LoS could move a galaxy into or out of a protocluster region, boosting or diminishing the measured overdensity, respectively. To account for this effect, we transform the LoS coordinate as follows:

$$d' = d + \frac{v_{\rm los}}{a(z)\, H(z)}\,. \tag{3.5}$$

Here d is the original comoving coordinate value, d' is the transformed coordinate, $v_{\rm los}$ is the peculiar galaxy velocity in the LoS direction, a is the expansion factor and $H(z)$ is the Hubble parameter at redshift z. This is derived as follows. The comoving distance, d_c, (assuming a homogeneous universe with a smooth expansion) is defined as

$$d_c = \frac{c}{a}\int_0^{\hat{z}} \frac{dz}{H(z)} \tag{3.6}$$

where $\hat{z}$ is the cosmological redshift, and $H(z)$ is the Hubble parameter. If a galaxy has a peculiar velocity component v_{pec} it will have an apparent distance measured from it's redshift $d_p \neq d_c$. This will be the result of the redshift due to the hubble flow, and the apparent redshift caused by the peculair velocity along the line of sight,

$$v = \hat{v} + v_{pec} \tag{3.7}$$

$$z = \frac{v}{c} \tag{3.8}$$

$$z \approx \hat{z} + z_{pec} \tag{3.9}$$

So we can rewrite the comoving distance as

$$d_c = \frac{c}{a}\left[\int_0^{\hat{z}} \frac{dz}{H(z)} + \int_{\hat{z}}^{\hat{z}+z_{pec}} \frac{dz}{H(z)}\right] \tag{3.10}$$

$$= \hat{d} + \frac{c}{a}\int_{\hat{z}}^{\hat{z}+z_{pec}} \frac{dz}{H(z)} \tag{3.11}$$

$$= \hat{d} + \frac{c}{a\,H_z}(z_{pec}) \tag{3.12}$$

$$= \hat{d} + \frac{v_{pec}}{a\,H(z)} \tag{3.13}$$

This equation applies in the comoving (fixed time hypersurface) frame. To correct observed peculiar motions one must use the proper distance, which leads to a multiple of the scale factor a in the distance equation,

$$d_p = c\int_0^{\hat{z}} \frac{dz}{H(z)} \tag{3.14}$$

which, when including peculiar velocities, gives the same equation for the apparent distance, minus the scale factor,

$$d_p = \hat{d}_p + \frac{v_{pec}}{H(z)}\ . \tag{3.15}$$

3.3 Results

3.3.1 The Protocluster Galaxy Population

We begin by looking at the evolution of the galaxy population as a whole from $2 \leq z \leq 9$ divided into protocluster and field designations. Figure 3.1 shows the Galaxy Stellar Mass Function (GSMF) for each selection criteria at each redshift, along with the biased GSMF for those galaxies that reside in protoclusters. The most massive galaxies are more likely to reside in protoclusters, and there is a dearth of low mass galaxies in protoclusters compared to the field, similar to trends seen in protocluster observations (Steidel et al., 2005; Strazzullo et al., 2013; Cooke et al., 2014). The normalisation is significantly lower at the intermediate to low mass range, similar to that seen in the $z < 1$ cluster environment

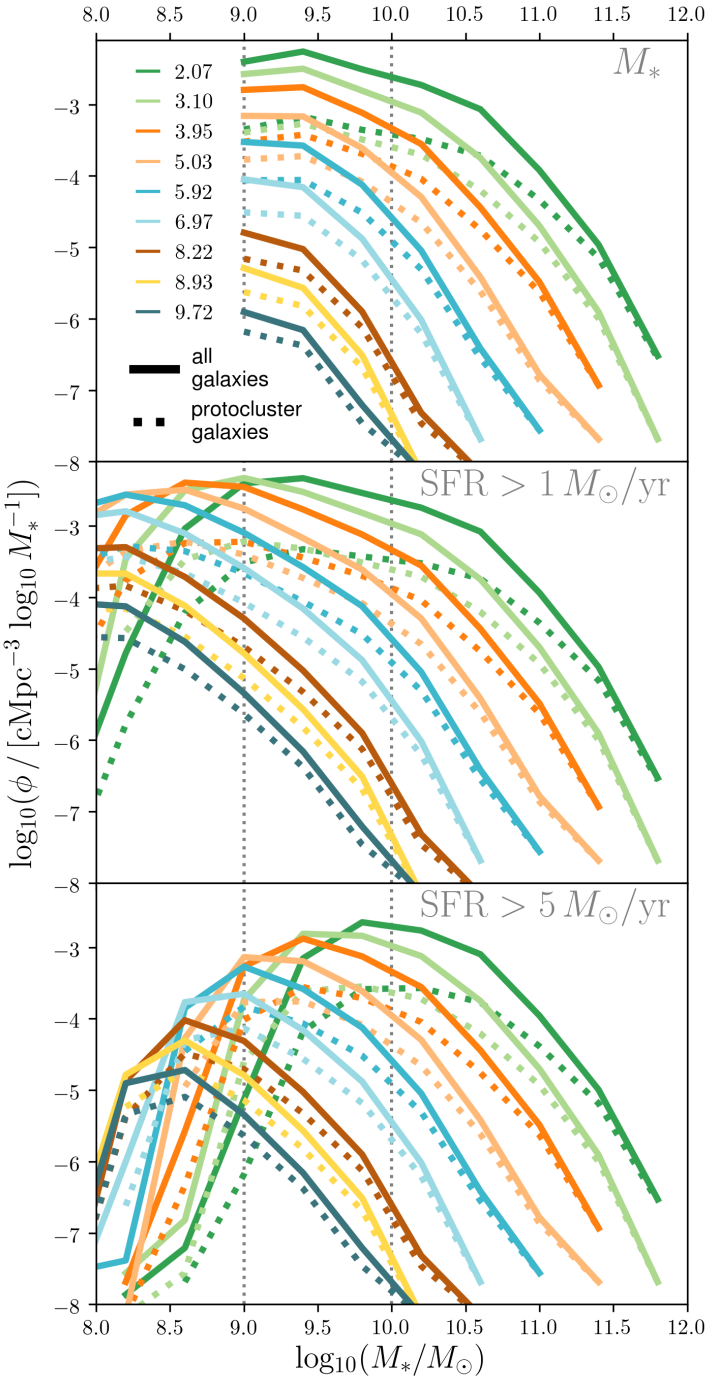

Figure 3.1: GSMF for all selections. The vertical dotted lines delimit the S_{MAS9} and S_{MAS10} selections. Solid lines show the full galaxy population, dashed lines show galaxies in protoclusters. The highest mass galaxies preferentially appear in protocluster environments, and there is a dearth of low mass galaxies evidenced by the flat low mass slope, as seen in Muldrew et al. (2015) for a previous version of the model. S_{SFR1} extends to lower stellar masses, but has little effect on the high mass end. S_{SFR5} truncates the selection of low mass galaxies, though the shape of the high mass slope is again unaffected.

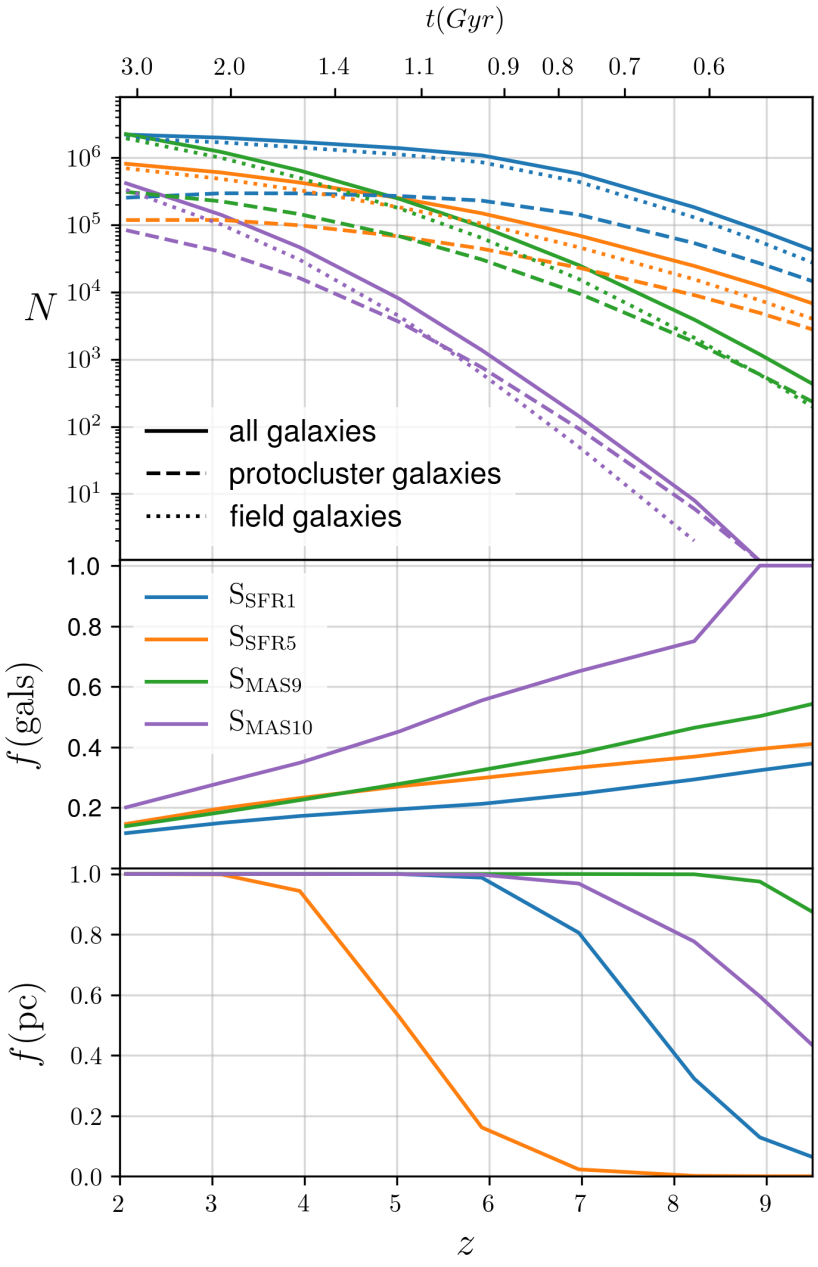

Figure 3.2: *Top:* Number of galaxies over time, for all galaxies (solid), protocluster galaxies (dashed) and field galaxies (dotted), for each selection. *Middle:* The fraction of galaxies in each selection that reside in protoclusters. *Bottom:* The fraction of protoclusters that contain at least one galaxy in the given selection.

(Vulcani et al., 2011).

The top panel of Figure 3.2 shows the number of galaxies over cosmic time, split into field and protocluster populations. The number of star forming (S_{SFR1} & S_{SFR5}) galaxies in protoclusters plateaus at $z \sim 5$, whilst similarly star forming galaxies continue to increase in number in the field. The middle panel shows the fraction of all galaxies from each selection that reside in protoclusters; at $z = 2$ a minority (10-20 %) of galaxies lie in protoclusters, rising to $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$ and 1 at $z > 9$ for S_{SFR1}, S_{SFR5}, S_{MAS9} and S_{MAS10}, respectively. Conversely, the bottom panel of Figure 3.2 shows the fraction of protoclusters that contain *at least one* galaxy from each selection; all protoclusters contain at least a S_{MAS9} mass galaxy up to the most extreme redshifts, whereas S_{SFR5} galaxies are only observed in a majority of protoclusters at $z < 5$.

For S_{MAS10} galaxies at $z > 6$ there is a $> 50\%$ chance they reside in a protocluster, and $> 50\%$ of all protoclusters contain at least one S_{MAS10} galaxy up to extreme redshifts; such galaxies can act as beacons of protocluster regions solely by virtue of their existence.

3.3.2 Triaxial Modelling

We have seen that protocluster galaxy membership evolves significantly with redshift and depends critically on the selection. We now look at the distribution of galaxies within protoclusters, and present the first model of protocluster shapes, a simple triaxial model of the galaxy spatial distribution at high redshift, in order to determine the extent to which they differ from the simplifying assumption of spherical symmetry. We acknowledge that such a simple model cannot probe collapsed structure such as groups and filaments within the protocluster, but it is capable of tracing the most prominent structure (if it exists), and provides insight into the global spatial asphericity, important for overdensity measurements.

The length and direction of each semi-axis in the triaxial model can be derived from the eigenvalues and eigenvectors, respectively, of the inertia tensor of the galaxy distribution. The components of the inertia tensor are given by

$$\boldsymbol{I}_{ij} = \sum_{n=1}^{N_g} (\boldsymbol{r}_n^2 \, \delta_{ij} - r_{n,i} r_{n,j}) \, ,$$

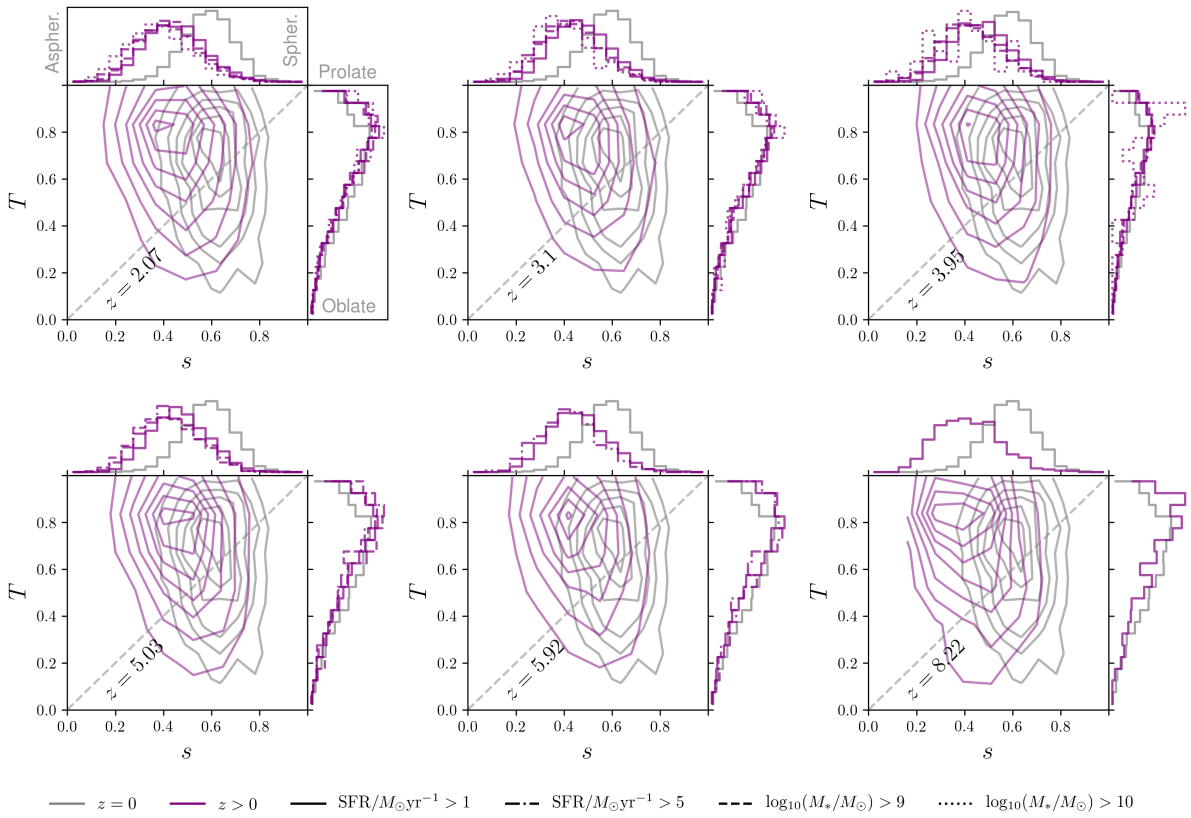

Figure 3.3: s ratio (a measure of sphericity) and T parameter (a measure of the form of apshericity) distributions. Each panel shows the 2D (for S_{SFR1}) and marginal (selection labelled) distributions at a given redshift. Values of s close to 1 indicate spherical distributions, values close to 0 aspherical. Values of T close to 1 indicate prolate distributions, values close to 0 oblate; if the s distribution suggests a spherical distribution then the nature of the asphericity is unimportant. Protoclusters tend to be aspherical, with a prolate distribution, and this asphericity is pronounced at high redshift. The $z = 0$ distributions (for S_{MAS9}, since there are an insufficient number of galaxies with high star formation rates at high-z) are shown in grey for comparison.

where N_g is the number of galaxies in the protocluster, $\boldsymbol{r}_n$ is the position vector of the n^{th} galaxy, and i and j are the tensor indices ($i, j \in 1, 2, 3$). We ignore the full matter distribution and focus on observable tracers, setting all galaxies to have equal mass, and also ignore redshift space distortions, so that any asphericity is randomly orientated. The moments of inertia of $\boldsymbol{I}$ are given by its eigenvalues, $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$, which can be translated into the relative axis lengths ($a \geqslant b \geqslant c$):

$$a = \sqrt{\frac{5}{2N_g}(\lambda_1 + \lambda_2 - \lambda_3)} \tag{3.16}$$

$$b = \sqrt{\frac{5}{2N_g}(\lambda_1 + \lambda_3 - \lambda_2)} \tag{3.17}$$

$$c = \sqrt{\frac{5}{2N_g}(\lambda_2 + \lambda_3 - \lambda_1)}\,, \tag{3.18}$$

Using these axis lengths we introduce three axis ratios,

$$s \equiv \frac{c}{a}, \quad q \equiv \frac{b}{a}, \quad p \equiv \frac{c}{b}\,. \tag{3.19}$$

Of these, s is of particular value as a measure of sphericity: where $s = 1$, the distribution is spherical, and where $s \sim 0$, the distribution is highly aspherical. The q and p ratios can be used together to deduce the form of the asphericity: where $q \sim 1(0)$ the distribution is oblate (prolate), and where $p = 1(0)$ the distribution is prolate (oblate). An alternative measure of the form of the asphericity is the Triaxiality parameter (Franx et al., 1991),

$$T = \frac{a^2 - b^2}{a^2 - c^2} \tag{3.20}$$

which measures whether an ellipsoid is prolate ($T = 1$) or oblate ($T = 0$), but does not measure the degree of asphericity.

Similar shape analysis has been applied to a range of astrophysical objects, including the profiles of cluster dark matter halos (Thomas et al., 1998; Wu et al., 2013). In such cases the reduced inertia tensor, which weights particles near the centre of the halo more highly, is often used (Schneider et al., 2012). Since protocluster profiles are less centrally concentrated than clusters (it is often difficult to unambiguously identify the protocluster centre), and are more likely to contain multiple filamentary structures, we

use the unweighted inertia tensor to characterise the entire shape. Bett et al. (2007) also note that particle discreteness can affect the determination of shape parameters using the inertia tensor; to mitigate this effect we ignore those selections where the average number of tracer galaxies in a protocluster drops below 20 at a given redshift.

Figure 3.3 shows the combined and marginal distributions of the s ratio and T parameter at different redshifts[5]. At $z = 0$ (shown in grey) the majority of clusters, as traced by their galaxies, are mildly aspherical with a prolate configuration.[6] Protoclusters, in comparison, are more aspherical, and the majority are prolate[7].

The S_{MAS9} and S_{MAS10} selections (shown in the marginal distributions of Figure 3.3) exhibit greater asphericity than those selected by star formation rate: those tracer galaxies that make the selection cut tend to be arranged along a single axis, leading to lower values of s. This suggests that care must be taken when using highly biased selections so as not to miss galaxies apherically distributed around the protocluster outskirts.

We see evidence in the evolution of s and T for the emergence of a red sequence. Between $z = 8.93$ and $z = 3.95$, $\bar{s}$ rises steadily from 0.36 to 0.49, then falls to 0.45 by $z = 2.07$. There is no dramatic collapse in spatial extent over this period which could explain the fall in s (Muldrew et al., 2015); most of the collapse to form current-day clusters occurs at $z < 2$. Instead, we attribute it to a decrease by a factor of 2 in the number of S_{SFR1} galaxies between $z = 2$ and 3, with the decrease predominantly toward the center of each protocluster (for which we see evidence in Figure 3.5): those galaxies that do make the S_{SFR1} cut are distributed irregularly outside the protocluster centre, leading to aspherical distributions.

3.3.3 Spherical Profiles

Galaxy overdensities are typically measured within cylindrical apertures along the line of sight (Franck & McGaugh, 2016a). Section 3.3.2 shows that protocluster galaxies tend to

[5]There is significant evolution in the number of galaxies in protoclusters selected by stellar mass or star formation rate throughout cosmic time, necessitating comparison between selections where there are insufficient galaxies to make a robust shape measurement: for example, galaxies at $z = 0$ are selected using the S_{MAS9} criteria, since there are not enough galaxies with high star formation rates at late times, and at $z \geq 2$ only S_{SFR1} is shown for the combined distribution, as it is the most populous selection.

[6]At $z = 0$, $\bar{s} = 0.61$, $\sigma_s = 0.10$, $\bar{T} = 0.61$, $\sigma_T = 0.20$. This asphericity is greater than that measured using the full dark matter particle information (Schneider et al., 2012).

[7]At $z = 3$, $\bar{s} = 0.50$, $\sigma_s = 0.12$, $\bar{T} = 0.65$ and $\sigma_T = 0.20$.

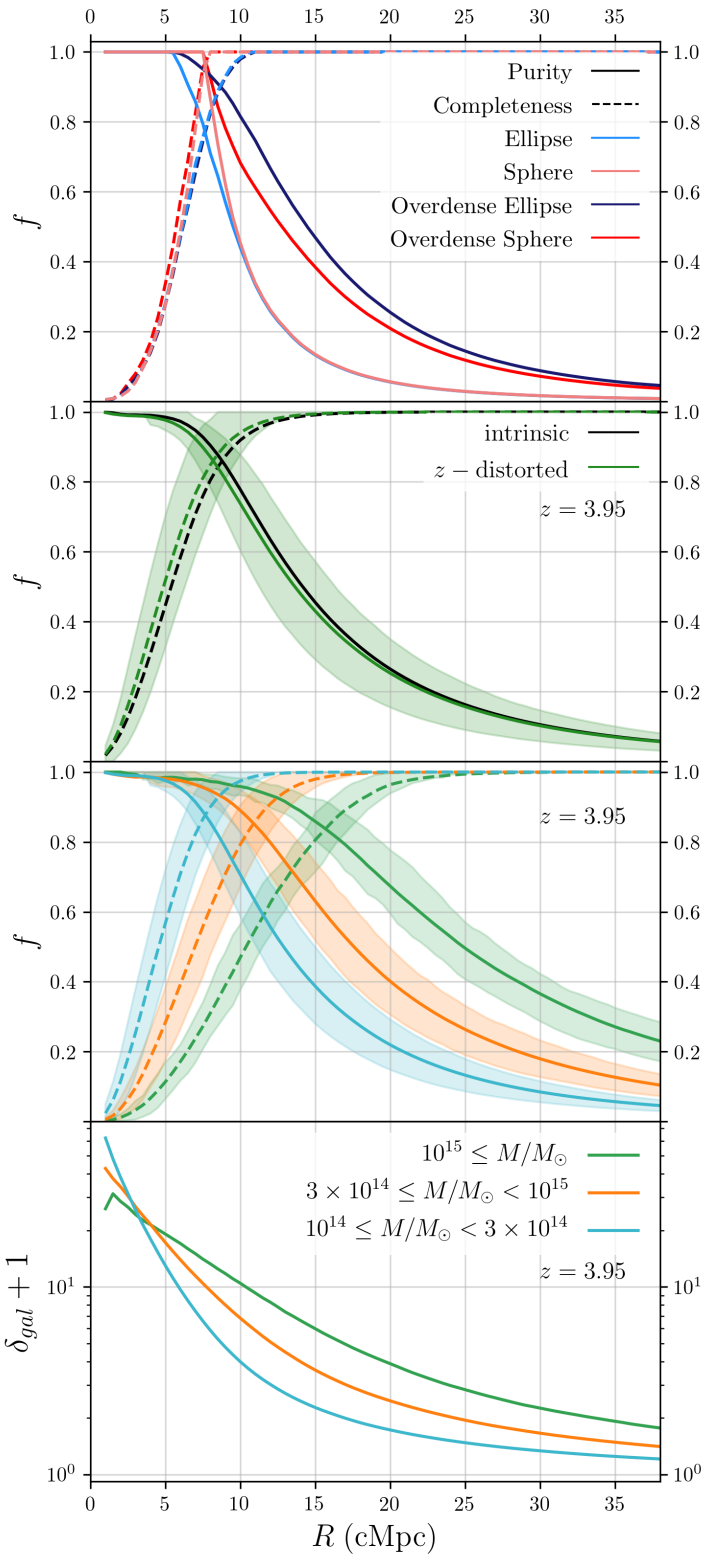

Figure 3.4: Average spherical profiles of protocluster galaxy properties in comoving coordinates. *Top panel:* Theoretical completeness (dashed) and purity (solid) profiles for a model ellipse and sphere with $\delta_g + 1 = 1$ and $\delta_g + 1 = 5$. *Second panel:* Mean purity and completeness profiles of the protocluster galaxy population at $z = 3.95$ for the S_{SFR1} selection. Intrinsic (black) and redshift space distorted (green) curves are shown, along with their 16th-84th percentile range. *Third panel:* The same redshift space distorted profile as in the second panel, split into three descendant cluster mass bins. *Bottom:* stacked galaxy overdensity profiles (including redshift space distortions), split into three descendant cluster mass bins.

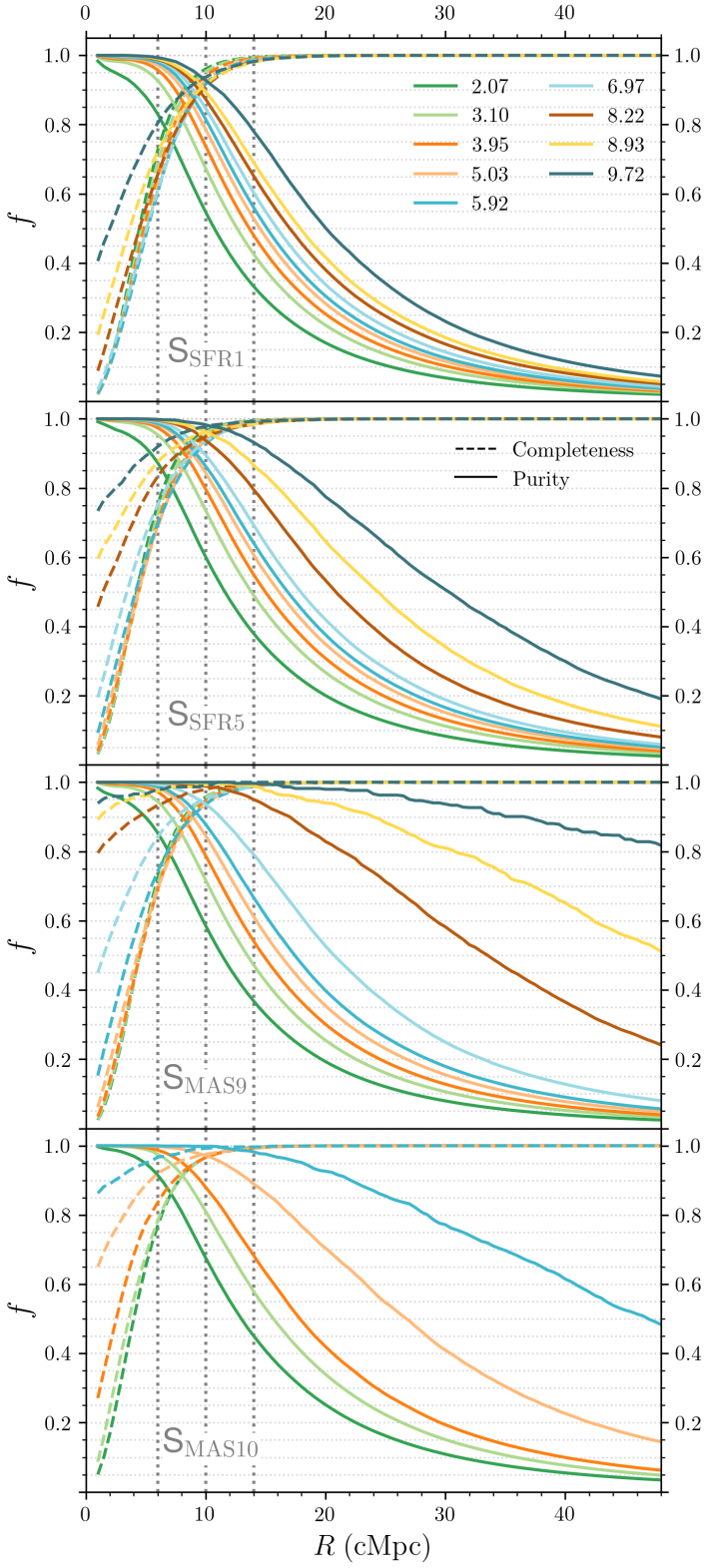

Figure 3.5: Mean completeness (dashed) and purity (solid) profiles for the protocluster population at a range of redshifts (labelled in the top panel). Panels top to bottom show the S_{SFR1}, S_{SFR5}, S_{MAS9} and S_{MAS10} selections, respectively. Vertical dashed lines show the approximate aperture sizes used in Figure 3.6.

be aspherically distributed with a prolate configuration, so such measurements could be biased by the introduction of many field galaxies, or by missing extended protocluster structure not contained within the aperture. To investigate we measure the properties of protoclusters as a function of radius from their centre (defined as the median coordinates of the selected protocluster galaxies), starting with the completeness and purity profiles of the galaxy population, before moving on to overdensity profiles.

3.3.3.1 Protocluster Galaxy Completeness and Purity Profiles

We begin by looking at the evolution in the completeness and purity of the protocluster galaxy population as a function of radius for a toy model ellipse. The volume of the ellipse represents the protocluster galaxy distribution, and outside represents the field. The shape of the model ellipse is based on the mean measured protocluster axis lengths for the S_{SFR1} selection at $z = 3.95$,[8], and initially assume the galaxy distribution is identical in both protocluster and field.

The purity and completeness as a function of radius can then be derived from the volume ratios, as shown in the top panel of Figure 3.4. The model ellipsoid is labelled 'Ellipse' and shown in blue, and a spherical model with the same volume is labelled 'Sphere' and shown in light pink. Close to the centre the completeness is low and the purity high, as expected; as the sphere is grown the completeness increases until it encapsulates all of the ellipse, whilst the purity begins to fall as more field volume is included. The curves cross at high values of both completeness and purity.

The second panel of figure 3.4 shows the mean completeness and purity curves for the protocluster galaxy population in L-GALAXIES at $z = 3.95$. We define the centre of the protocluster as the median of the protocluster galaxy coordinates, the completeness as the fraction of all protocluster galaxies within the aperture, and the purity as the fraction of galaxies within the aperture that are members of the protocluster. Both intrinsic (black) and redshift space distorted (green) coordinates are shown. The 16th-84th percentile range is shown as a shaded region; the majority of protoclusters exhibit similar profiles, and cross over at high values within a tight range of radii.

The purity and completeness curves both show gradual evolution toward the edge of the

[8] $a = 11.00$, $b = 7.56$ and $c = 5.36$ (cMpc)

protocluster, rather than the sudden change seen in our toy model, and the purity curve drops off much more gradually, which we attribute to our naive assumption of a uniform density of galaxies in our toy model – in reality, protoclusters have a higher overdensity than the surrounding field. To model this, we increase the number of samples in the ellipse by a factor of 5, simulating a galaxy overdensity of $\delta + 1 \sim 5$. The completeness and purity curves for this model are shown in the top panel of Figure 3.4 in dark blue, labelled 'Overdense Ellipse'; the purity curve falls much more gradually, as seen in the SAM. Importantly for measurements of galaxy overdensity, the lower number density of galaxies in the field acts to reduce the effect of asphericity on the measured galaxy population, lowering the contamination on the protocluster outskirts and maintaining relatively high purity out to large radii. It is not unreasonable then, when producing descriptive statistics on the protocluster population, to adopt spherical symmetry above some minimum radius.

The inclusion of redshift space distortions has two effects. The coherent motions of galaxies as they fall toward the centre of the forming cluster leads to an apparent flattening in their appearance, known as the Kaiser effect (Kaiser, 1987), and we see evidence for it in the steeper completeness curve; galaxies appear closer to the centre, which can be explained if they are, on average, infalling, (Contini et al., 2016), and this acts to marginally boost the overdensity measurement. The purity curve drops at lower radii, which suggests greater apparent contamination from field galaxies; these galaxies are gravitationally disturbed by the forming protocluster, but do not enter the virial radius by $z = 0$. The two curves still cross at high values ($> 80\%$).

The third panel of Figure 3.4 shows the mean completeness and purity curves for protoclusters at $z = 3.95$ split by descendant cluster mass. There is a positive correlation between cluster size and crossover radius: protoclusters with the most massive descendants trace larger volumes than those that will form lower mass clusters. In order to capture the majority of the galaxies in the most massive protoclusters a much larger field of view is required. However, the majority of protoclusters can be captured in their entirety using a much smaller aperture, and even the largest protoclusters contain a significant fraction of their tracer galaxies within a smaller aperture ($> 50\%$ at $R = 10$ cMpc). The crossover values remain high ($> 80\%$) for all mass bins.

Figure 3.5 shows the mean completeness and purity for each selection criteria with redshift. For the most stringent selections at the highest redshifts the completeness curves start at non-zero values, since some protoclusters may be represented by only a single galaxy, boosting the mean. Similarly, the purity curves also remain high, since where galaxies are rare in protoclusters, they also tend to be rare in the field; where they exist, they are highly clustered and located in protoclusters (see Figure 3.2). The purity curve falls at lower radii with decreasing redshift for all selections, caused by the protocluster collapse and central concentration, and the higher relative density of field galaxies with decreasing redshift (see Figure 3.2).

The exception to this evolution is seen at low redshift ($z \leqslant 3$) for both $\mathrm{S}_{\mathrm{SFR1}}$ and $\mathrm{S}_{\mathrm{SFR5}}$: the purity falls significantly at much lower R, and the completeness curve is also steeper. Figure 3.2 shows that the number of $\mathrm{S}_{\mathrm{SFR1}}$ protocluster galaxies decreases below $z = 3.10$, which can be explained by the emergence of a red sequence; since there are fewer star forming galaxies at the centre of protoclusters relative to the outskirts, the completeness curve rises more rapidly with radius. We see further evidence for the emergence of a red sequence in the asphericity distribution between $z = 3$ and 2 (see Section 3.3.2).

The crossover between purity and completeness remains high, $\geqslant 80\,\%$, and is relatively insensitive to changes in redshift or selection criteria. The cross over radii also all fall within a narrow range of values, which suggests a characteristic scale can be chosen,

$$R_{\mathrm{C}} \sim 10\,\mathrm{cMpc}\,, \tag{3.21}$$

that maximises the completeness and purity regardless of selection criteria or redshift. This corresponds approximately to an angular scale ($2R_{\mathrm{C}}$) of 10 arcmin on the sky at $z = 2$, falling to 6 arcmin by $z = 9$, not much larger than typical focused searches around biased tracers such as AGN.

3.3.3.2 Protocluster Galaxy Overdensity Profiles

The bottom panel of Figure 3.4 shows the differential stacked overdensity profiles, measured using all galaxies (protocluster+field) within a spherical aperture centred on the protocluster, and split by descendant mass. We find similar centrally peaked profiles to the surface overdensities measured in Overzier et al. (2009) & Chiang et al.

(2013). The slope of the overdensity profile at small-intermediate radii is shallower for higher mass protoclusters – they are less centrally concentrated and more extended – and for lower mass protoclusters they are more sharply peaked toward the centre. This may be as a result of our protocluster centre definition: lower mass protoclusters typically have only a single dominant group, so the centre will be defined within this group, leading to a peaked profile at low R. Conversely, in larger protoclusters with multiple similarly sized subgroups the median coordinates may lie in an intergroup region, lowering the measured overdensity on small scales. However, measuring the overdensity centred on a single subgroup will not be representative of the entire protocluster, and may lead to lower purity and completeness at larger radii. We therefore emphasise the need to make descendant mass estimates from overdensity measurements over sufficiently large apertures ($R > 7\,\mathrm{cMpc}$), which we demonstrate in Section 3.3.4.2. The variation in slope of the overdensity profile with mass suggests that measuring overdensity on multiple scales could lead to a more accurate descendant mass estimate, however we found that the improvement in the fit is not substantial.

3.3.4 Galaxy Overdensity Statistics

Protoclusters have irregular shapes, but this has a small effect on the completeness and purity of their galaxy populations when measured in a sufficiently large aperture. However, the size and shape of the aperture used to measure the overdensity can have a significant effect on the qualitative value of the overdensity (see the bottom panel of Figure 3.4, and Shattow et al. (2013)), on which further properties, such as protocluster probabilities and descendant masses, are based. We propose an improved procedure for deriving overdensities that takes into account irregular apertures.

3.3.4.1 Identifying Protoclusters in Galaxy Overdensities

We select 100 000 random regions, with surface area, $\pi\,R^2$, and depth, $D \equiv \Delta d'$, in the Millennium volume. We call each of these regions a *candidate*. For each galaxy in the candidate we find its descendant halo mass. If no galaxies in the candidate have cluster descendants, the candidate is labelled a field region. If there are cluster progenitors in the candidate, the completeness, C, and purity, P, of the galaxy population in this candidate with respect to each descendant cluster is calculated. Each region can then be classified as

Table 3.1: Candidate region labelling conditions. C is completeness, P purity, and $C_{\rm lim}$ and $P_{\rm lim}$ are limiting values of each that differentiate each classification.

Label	Condition	Description
Proto	$C \geq C_{\rm lim}$ and $P \geq P_{\rm lim}$	Protocluster region.
ProtoField	$C \geq C_{\rm lim}$ and $P < P_{\rm lim}$	Region traces the combination of a proto-cluster and field region.
PartProto	$C < C_{\rm lim}$ and $P \geq P_{\rm lim}$	Region traces a part of a protocluster.
Field	$C < C_{\rm lim}$ and $P < P_{\rm lim}$	Field region.

.

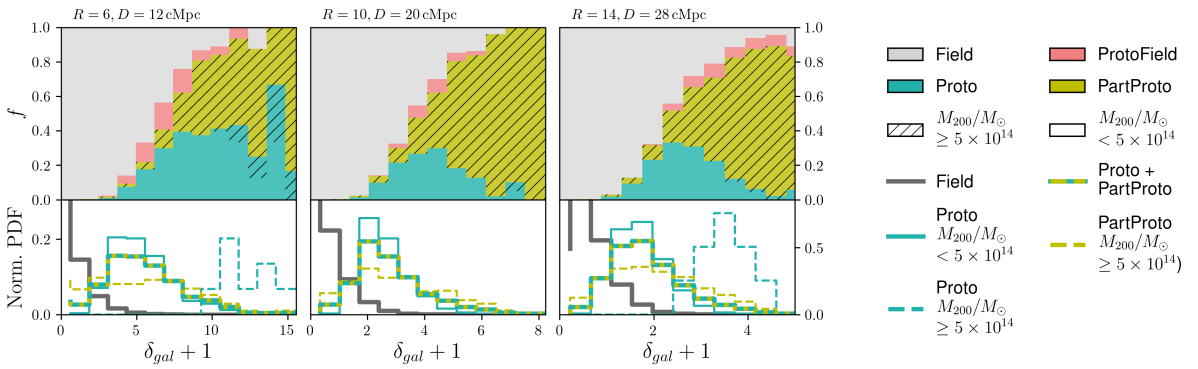

Figure 3.6: *Top:* Fractional probability distribution of candidate being Proto, PartProto, ProtoField or Field ($S_{\rm SFR1}$, $z = 3.95$). Where the distribution is hatched represents those candidates that trace high mass ($M_{200}/M_\odot \geq 5 \times 10^{14}$) protoclusters. Each panel shows a different aperture size, labelled at the top. We choose $C_{\rm lim}$ and $P_{\rm lim}$ values equal to the 5$^{\rm th}$ percentile of the completeness and purity of the protocluster population (for this aperture and selection). *Bottom:* Normalised probability density distributions for each classification, split into low and high mass descendants.

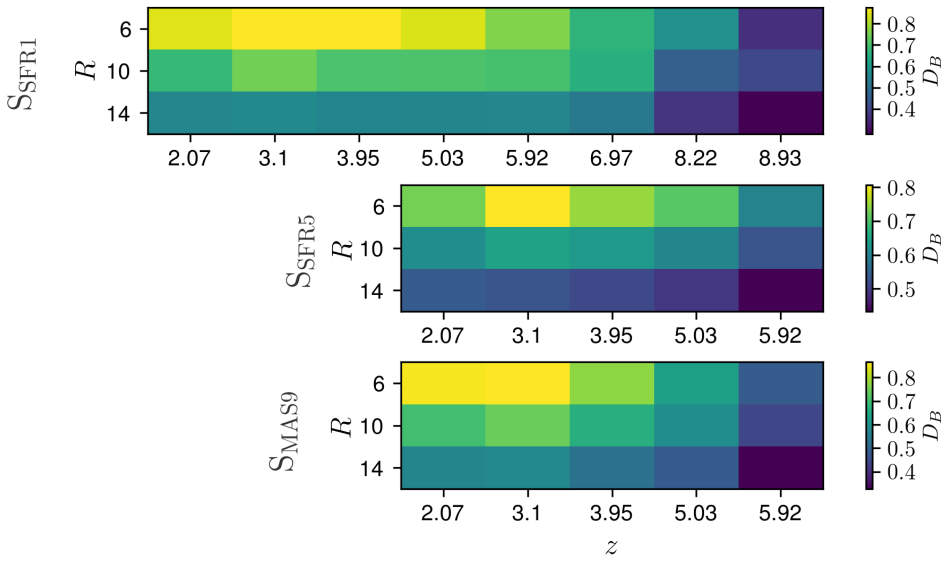

Figure 3.7: Colour map showing the Bhattacharrya distance (D_B) between the combined Proto+PartProto and Field distributions for the S_{SFR1} S_{SFR5} and S_{MAS9} selections, over a range of redshifts (z) and aperture sizes ($R = D/2$, cMpc). The S_{MAS10} selection, and some redshifts, are not shown since there are insufficient galaxies to produce a reasonable statistic. D_B is maximised at $R = 6$ for all selections at almost all redshifts, and decreases as the selection region is increased in volume.

Proto: 'protocluster', ProtoField: 'protocluster+field', PartProto: 'part of a protocluster', or Field: 'field' according to the conditions detailed in Table 3.1. In the rare case where there are multiple cluster descendants, the cluster with the highest value of the purity and completeness added in quadrature is chosen.

Importantly, the values of $C_{\rm lim}$ and $P_{\rm lim}$ are chosen based on the $5^{\rm th}$ percentile of the completeness and purity of the protocluster population *given the chosen selection criteria and aperture*. This allows a more accurate characterisation of candidate regions that takes into account the actual galaxy membership of protoclusters. For example, one would not expect to have high purity in a large aperture due to contamination from field galaxies on the outskirts, but would demand high completeness since the majority of a protoclusters galaxies should be captured. We demonstrate the effect of changing $C_{\rm lim}$ and $P_{\rm lim}$ whilst maintaining a fixed aperture in Appendix 3.6.1.

Once all candidates are labelled, we can calculate the fractional probability that a measured overdensity represents one of our 4 labels, further split by the mass of the descendant cluster. Figure 3.6 shows an example; the upper panel shows the fractional probability distribution, the lower panel the probability density distribution. The default parameters

are $R = D/2 = 10\,\mathrm{cMpc}$ and $z = 3.95$, using the $\mathrm{S_{SFR1}}$ selection, and we choose C_{lim} and P_{lim} values equal to the 5^{th} percentile of the completeness and purity of the protocluster population with this aperture and selection. As expected, higher galaxy overdensities are more likely to evolve into clusters, and the highest overdensities are more likely to form more massive protoclusters. At intermediate to high overdensities, a considerable fraction of candidates trace PartProto regions. All of these PartProto candidates trace high mass protoclusters; lower mass protoclusters cannot satisfy C_{lim} whilst simultaneously satisfying P_{lim} as they are not large enough. At intermediate overdensities there is a small probability that a candidate is probing a ProtoField region, and these are all for smaller, lower mass protoclusters.

The approach is similar to that demonstrated in Chiang et al. (2013), though the criterion for classifying a random region as a protocluster is different: they require that the center of the random region lies within half a box length of a protocluster centre, so that the window covers, on average, $> 50\%$ of the protocluster mass.[9] Our analysis in Section 3.3.2 and Section 3.3.3 suggests that the assumption of spherical symmetry is violated, particularly at high-z, so this definition may identify regions with significant field galaxy populations. Despite these differences (including the use of an updated version of L-Galaxies and the Planck cosmology) we achieve consistent results: the protocluster fractions of $\mathrm{S_{SFR1}}$ galaxies at $z \sim 4$ match the combined Proto and PartProto distribution in the right panel of Figure 3.6, with a slight shift in quantitative overdensity to lower values (possibly due to using a slightly larger volume). The probability density distribution for low mass protoclusters appears to show less distinction from the field distribution as seen in Figure 6 in Chiang et al. (2013), which may be attributed to the updated protocluster definition, or to the change in cosmology.[10] Whilst consistent, we note that our approach explicitly distinguishes protoclusters identified partially or in whole, and can handle irregularly shaped apertures.

The probability density distributions at the bottom of each panel can be used to evaluate the separation in overdensity space of protocluster and field regions. We determine the Bhattacharyya distance (Bhattacharyya, 1946), a measure of the dissimilarity between

[9] private correspondence

[10] The Planck cosmology used in Henriques et al. (2015) leads to an increased dark matter particle mass, an increased box size, and the $z = 0.12$ output of the original WMAP1 simulation becomes the new $z = 0$; the latter two effects would lead to a diluted quantitative overdensity measurement

Table 3.2: Protocluster mass estimate fit parameters for Equation 3.23, for the S_{SFR1}, S_{SFR5} and S_{MAS9} selections, with error estimates.

Selection	a	b	c	C	R^2
S_{SFR1}	0.146	-1.077	2.628	1.752	0.547
S_{SFR5}	0.658	-1.317	1.859	0.117	0.549
S_{MAS9}	2.883	-1.681	1.452	-0.235	0.507

two probability distributions, defined as

$$D_B = -\ln BC, \quad \text{where } BC(p,q) = \sum_{\delta \in \Gamma} \sqrt{p(\delta)q(\delta)} \tag{3.22}$$

and p and q are the probability distributions over the galaxy overdensity domain Γ. D_B, calculated between the Field and combined Proto and PartProto distributions for a range of redshifts, aperture sizes and selections, is shown in Figure 3.7. At lower redshifts the distinction is greatest on small scales ($R = 6$ cMpc) for all selections, though the distinction on the characteristic scale ($R = R_C = 10$ cMpc) is still relatively high compared to larger scales. At higher redshifts the distinction is greatest at R_C. This seems to suggest that, in order to best separate protoclusters from the field, one should use a smaller aperture at lower redshifts and a slightly larger one at higher redshifts. However, the overdensity profiles shown in Figure 3.4 show that a larger aperture allows the greatest discrimination of protocluster descendant mass, and in Section 3.3.5 we find that, in searches surrounding AGN, D_B is maximised at R_C due to the non-central location of the AGN within the protocluster. We therefore still recommend making overdensity measurement on a scale of R_C for all redshifts and selections.

3.3.4.2 Protocluster Mass from Galaxy Overdensity

We now explore the relationship between high redshift overdensity and descendant cluster mass by fitting an empirical relation between the two. We fit to all halos at $z = 0$ with masses $M_{200}/M_\odot > 10^{13}$ in order to fully assess the spread in descendant masses for a given overdensity, calculating the overdensity measured in a single cylindrical aperture with radius and depth equal to the characteristic scale, $R_C = 10$ cMpc; on smaller scales descendant mass cannot easily be distinguished through galaxy overdensity (see Figure 3.4, bottom panel).

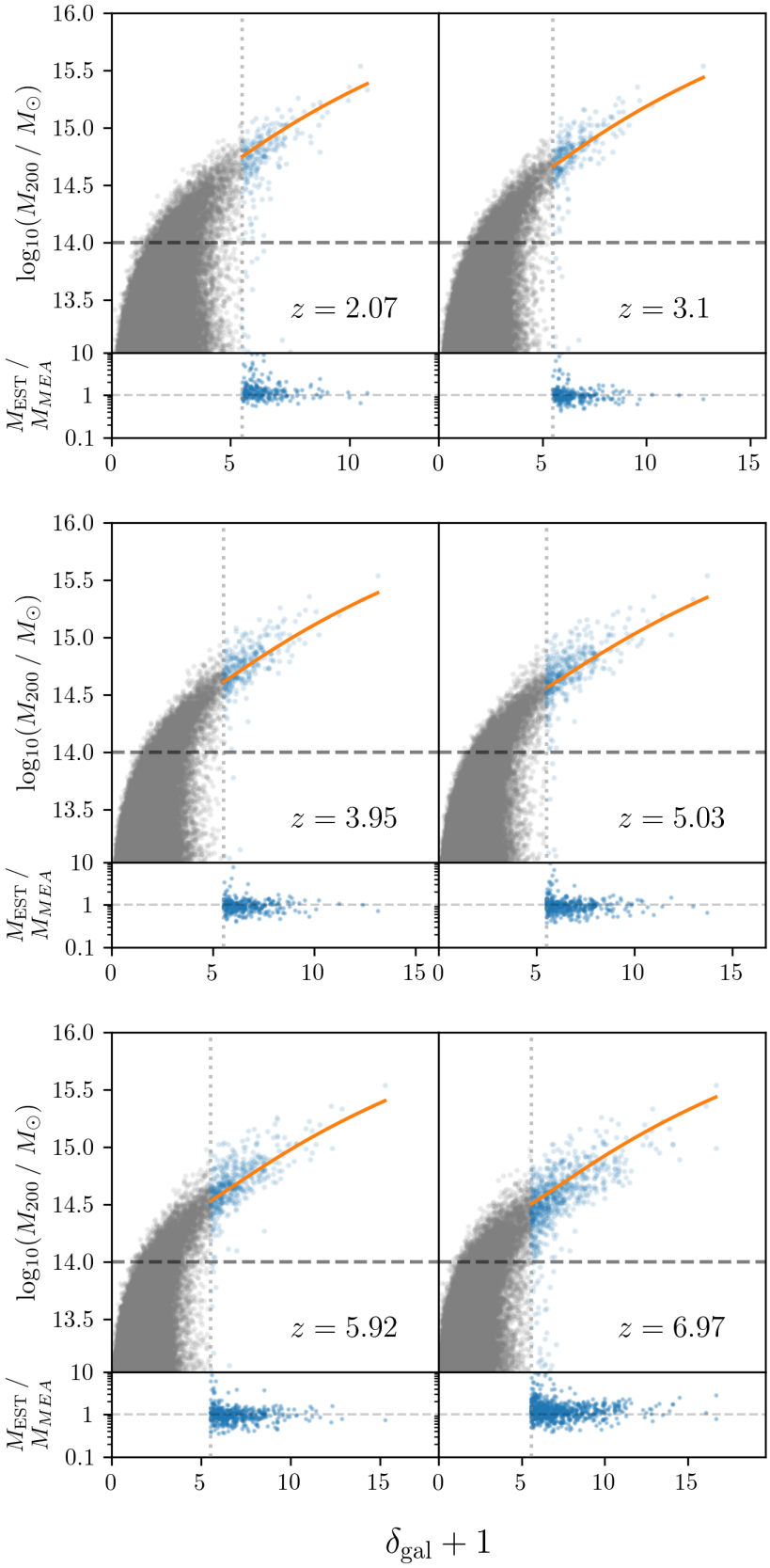

Figure 3.8: *Top panels:* Galaxy overdensity (S_{SFR1}) against descendant halo mass for all halos with $\log_{10}(M_{200} / M_\odot) > 13$. The fit at each redshift is shown in orange. Those objects used in the fit are shown in blue, those below the overdensity threshold in grey. Our cluster mass definition ($\log_{10}(M_{200} / M_\odot) > 14$) is delimited by the horizontal dashed black line. *Bottom panels:* Ratio of the estimated and measured masses.

The relationship between overdensity and descendant mass is parameterised as follows:

$$M_{200}/(10^{14} M_{\odot}) = a\,(1+z)^{b}\,(1+\delta)^{c} + C\,. \tag{3.23}$$

where M_{200} is the descendant mass, and δ is the measured galaxy overdensity. We fit the S_{SFR1}, S_{SFR5} and S_{MAS9} distributions between $z = 2-7$ using the `curve_fit` least squares minimisation method provided by SCIPY (Jones et al., 2001). The fit is illustrated in Figure 3.8 for the S_{SFR1} selection, with residuals shown at the bottom of each panel. We ignore both the S_{MAS10} selection and $z > 7$ due to a lack of galaxies. A striking feature of Figure 3.8 is the spread in descendant halo masses for $\delta_{gal} < 4.5$. We cannot make any meaningful descendant mass prediction below this overdensity limit, so we limit our fit to above this range; whilst there is a chance that such regions do trace protoclusters, the vast majority of them do not. The exact choice of threshold overdensity depends on many factors that affect the overdensity distribution (aperture size, selection, etc.). For this aperture, the distribution conveniently turns over at descendant masses of $\sim 10^{14}\,M_{\odot}$, which makes distinguishing high mass protoclusters by overdensity somewhat easier; lower mass protoclusters are harder to distinguish from the field.

A non-linear relationship provides a marginally better fit for the very highest descendant masses. In Section 3.3.3 we noted that the shape of protocluster overdensity profiles was dependent on their descendant mass, but including overdensity measured on two scales leads to no appreciable improvement in the fit, which we attribute to the scatter in overdensity profiles.

Chiang et al. (2013) derive a similar relation between overdensity and descendant mass, ignoring redshift space distortions, but taking into account the aperture size, whilst the coefficients of our empirical model must be rederived for differing apertures. We note that they only apply it to overdensities surrounding protoclusters, which underestimates the scatter in descendant halo mass at intermediate overdensities (see Figure 3.8), and in their Figure 12 showing the residuals they ignore objects with descendant masses below the protocluster mass threshold.

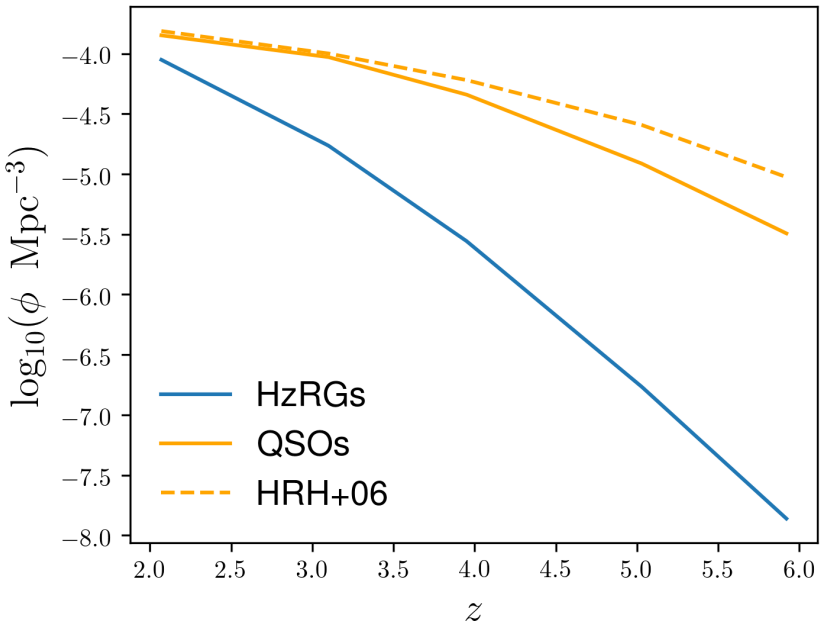

Figure 3.9: Number density evolution of HzRGs (blue) and quasars (solid orange) subject to the accretion cuts stated in Section 3.3.5. The quasar mode accretion cut was selected in order to match the number density evolution as measured by Hopkins et al. (2007) (dotted orange).

3.3.5 AGN as Protocluster Tracers

Both quasars and High Redshift Radio Galaxies (HzRGs) are expected to act as tracers of protocluster regions. In order to test this assumption we select a sample of quasars and HzRGs whose number densities match observations at high-z (Section 3.3.5.1), find their surrounding galaxy overdensities (Section 3.3.5.2) and investigate their coincidence with protoclusters (Section 3.3.5.3).

3.3.5.1 AGN selection

We choose our quasar mode accretion cut in order to match the integrated number densities from Hopkins et al. (2007) between $z = 2 - 5$ (assuming a lower luminosity limit of $10^{44}\ L_{\mathrm{bol}}\,/\,\mathrm{erg\ s^{-1}}$):

$$\dot{M}_{\bullet}(\mathrm{quasar})/(M_{\odot}\,\mathrm{yr}^{-1}) > 0.0036\ . \tag{3.24}$$

This gives a reasonably good fit to the normalisation and redshift evolution (see Figure 3.9). The accretion rate can be translated into a bolometric luminosity through the following prescription,

$$L_{\mathrm{bol}} = \epsilon \dot{M}_{\bullet} c^2 \tag{3.25}$$

where $\dot{M}_\bullet$ is the accretion rate and $\epsilon = 0.1$. For the quasar accretion mode this gives a lower limit of $L_{\rm bol} > 2 \times 10^{43}$ ergs s^{-1}, somewhat lower than typical intermediate-luminosity quasars, which suggests an underprediction of the black hole accretion rate at high-z.

Figure 3.9 shows a similar decline in number density of HzRGs in the model from $z \sim 2$. There is still significant uncertainty about the position and luminosity dependence of a high redshift cutoff in observations (Jarvis et al., 2001; Venemans et al., 2007; Rigby et al., 2011); we therefore choose a radio mode accretion threshold in order to approximately match the number densities measured by Dunlop & Peacock (1990) for the most powerful radio galaxies over the redshift range $z = 2 - 5$:

$$\dot{M}_\bullet({\rm radio})/(M_\odot\,{\rm yr}^{-1}) > 0.001\ . \tag{3.26}$$

We also adopt more conservative accretion cuts in order to test any dependence on the chosen cut-off (see Section 3.3.5.2).

Each panel of Figure 3.10 shows the distribution of black hole accretion rates as a function of host halo mass, for a range of redshifts, along with the marginal distribution of halo masses for the total AGN population and our selection. Each accretion mode is distinct: HzRG tend to populate higher mass halos, with a median mass $\log_{10}(M\,/\,M_\odot) \sim 12.5$, as expected since it is only the most massive halos that have a sufficient reservoir of hot gas to power this accretion mode. Quasars populate a much wider range of halo masses with a lower median mass of $\log_{10}(M\,/\,M_\odot) \sim 11.5$ at all redshifts considered. The quasar mode accretion rate is proportional to the product of the ratio of the masses of the merging galaxies and their combined cold gas mass, $\dot{M}_\bullet({\rm quasar}) \propto M_{\rm sat}\,/\,M_{\rm cen} \times M_{\rm cold}$. Whilst major mergers of high mass halos are rare, high quasar mode accretion rates can still be achieved in massive halos through minor mergers where the primary halo has a large gas reservoir.

3.3.5.2 Galaxy Overdensities Surrounding AGN

Given our AGN selection criteria from Section 3.3.5.1, figures 3.11 and 3.12 show the galaxy overdensity ($S_{\rm MAS9}$) in the vicinity of each quasar and HzRG (respectively) against its descendant halo mass for a range of redshifts and aperture sizes.[11] Each coloured

[11] for brevity we use regular apertures, $R = D/2$.

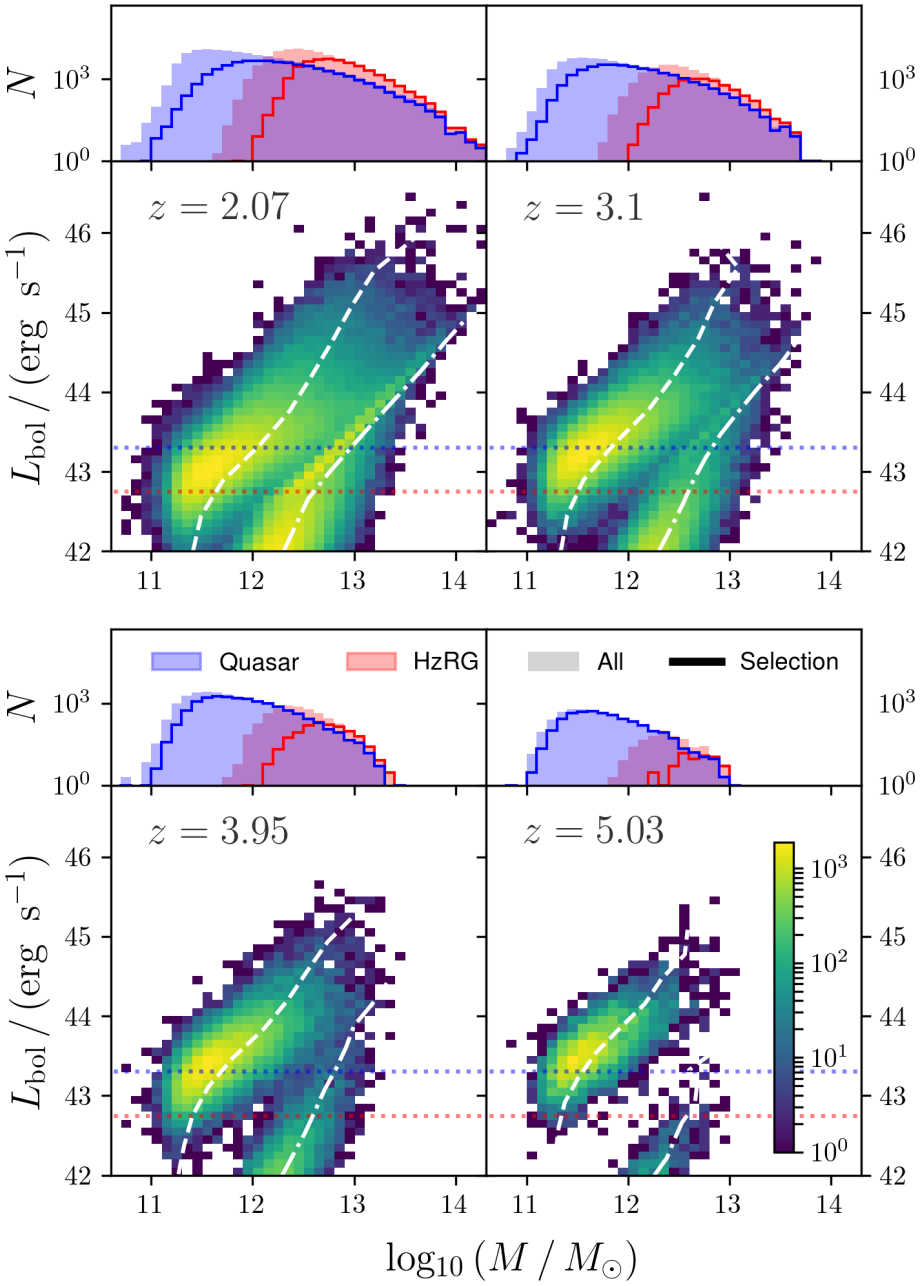

Figure 3.10: Distribution of AGN luminosity against host halo mass, for a range of redshifts. *Bottom panels:* 2D distribution of bolometric luminosity for the combined radio & quasar accretion modes against host halo mass. White dashed and dash-dot lines show the independent median of the relationship for the *quasar* and *radio* accretion modes, respectively. Horizontal red and blue dashed lines delimit the accretion cuts stated in Section 3.3.5. *Top panels:* Marginal distribution of host halo masses for the whole AGN population as filled histograms, and as step histograms for the accretion cuts stated in Section 3.3.5.

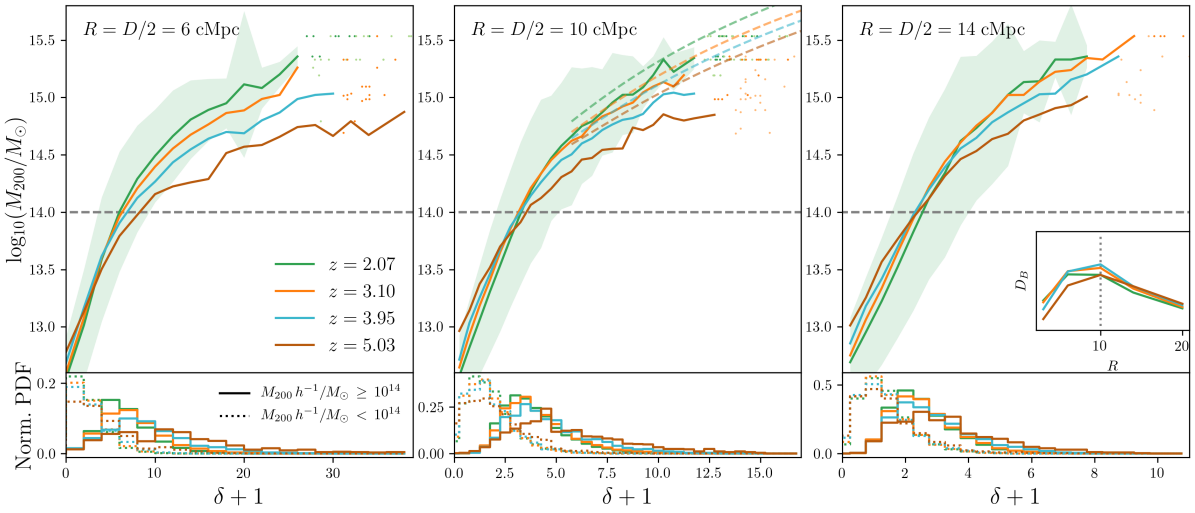

Figure 3.11: *Top:* Galaxy overdensity (S_{MAS9}) in the vicinity of quasars (selected according to the criteria in Section 3.3.5) against descendant halo mass. Solid lines show the binned mean, and the shaded region shows the 16th-84th percentile range for the $z = 2$ selection. Where there are less than 20 quasars in a bin, individual objects are plotted. The fit from Section 3.3.4.2 is shown as the dashed line in the central panel. *Bottom:* Probability density functions (PDF) for those quasars that evolve into clusters, and those that do not. *Inset:* Bhattacharrya distance, D_B, between the PDF for quasars that evolve into clusters and those that do not, as a function of aperture size. The peak indicates the aperture size at which AGN embedded in protoclusters are best discriminated from the field.

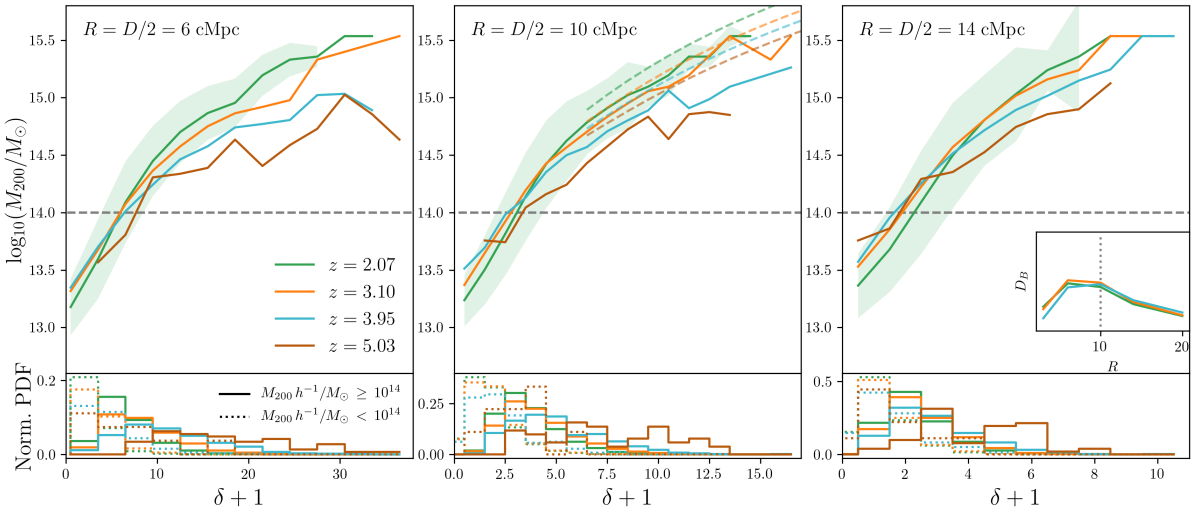

Figure 3.12: As for Figure 3.11, but for the HzRG selection.

line shows the binned mean for all AGN at each redshift, and $16^{\text{th}} - 84^{\text{th}}$ percentiles are shaded for the $z = 2$ selection. These figures can be used to read off the estimated descendant halo mass of an AGN given its surrounding galaxy overdensity.

The bottom of each panel shows the normalised probability density distribution for those AGN that end up in clusters and those that do not, in solid and dotted lines respectively, which can be used to calculate the Bhattacharrya distance (introduced in Section 3.3.4.1) to evaluate their level of separation in overdensity space. D_B is shown as a function of R in the inset figure in the third panel of each figure; it peaks between $5 - 10$ cMpc for both selection, but slightly higher for quasars. This is also higher than that seen for random regions of the same size in Section 3.3.4.1; this can be explained by the non-central location of AGN within protoclusters. For protoclusters containing quasars, the median distance of the quasar from the centre is $\sim 5.05\,$cMpc at $z = 3.95$; apertures of size ~ 10 cMpc capture the greatest proportion of the overdense protocluster whilst minimising field contamination, boosting the overdensity associated with that AGN, whereas smaller apertures sample the low overdensity tail. For HzRGs we see a similar trend with radius, but D_B peaks at lower radii, which can be attributed to the fact that the median distance of HzRGs from the centre of their host protocluster is lower (3.04 cMpc at $z = 3.95$). Hatch et al. (2014) find that radio loud AGN appear to reside in average overdensities on scales of 0.5 Mpc, but overdense environments on larger scales, in agreement with this interpretation.

The location of each AGN type within protoclusters can be explained by their differing treatment in the model. HzRGs preferentially appear in higher mass halos; during cluster assembly a dominant subhalo, with mass $M/M_{\odot} \sim 10^{12}$ emerges at intermediate redshifts (Chiang et al., 2013), typical of HzRG hosting halos, and will either already be at the center of the protocluster region or will migrate towards it. In contrast, high luminosity quasars can be triggered by both major and minor merger activity; whilst there will be many minor mergers with massive halos in the dominant subhalo, there will also be a large number of major mergers between intermediate mass halos elsewhere in the protocluster, so that the average quasar location is further from the protocluster centre.

The mass predictions from Section 3.3.4.2 are shown as dashed lines in the centre panel. Puzzlingly, the predicted descendant mass for a given overdensity is lower for AGN

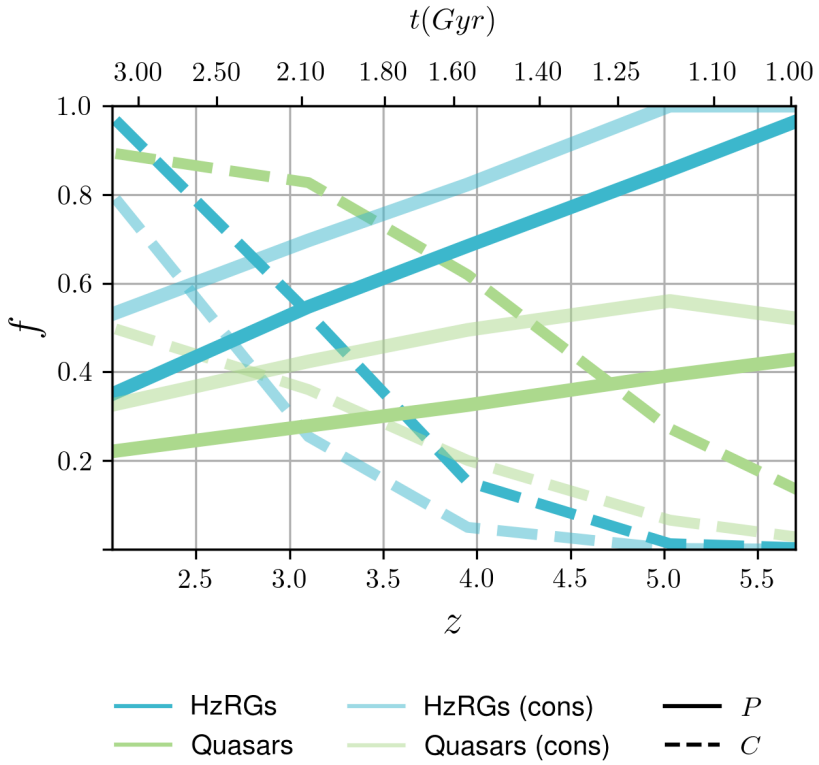

Figure 3.13: The completeness (dashed), and purity (solid) of AGN as protocluster tracers, for both HzRGs (blue) and quasars (green), and for both accretion thresholds (see Section 3.3.5.1 and Section 3.3.5.3).

than protoclusters: one would expect, for a given protocluster, the centrally measured overdensity to be larger than from the non-central AGN. We attribute this to a selection effect; not all protoclusters contain AGN at these redshifts, so the selection does not necessarily have the same descendant mass distribution.

3.3.5.3 The Coincidence of AGN & Protoclusters

Figure 3.13 shows the *completeness* and *purity* of AGN as biased tracers of protoclusters, where *completeness* in this context refers to the fraction of all protoclusters traced by AGN, and *purity* to the ratio of protoclusters to field regions traced. In order to assess the effect of our accretion cut choice, we also show the following more conservative accretion cuts:

$$\dot{M}_\bullet(\text{radio})/(M_\odot\,\text{yr}^{-1}) > 0.004 \tag{3.27}$$

$$\dot{M}_\bullet(\text{quasar})/(M_\odot\,\text{yr}^{-1}) > 0.018\,. \tag{3.28}$$

For both selections, at low redshifts the completeness tends to be high and purity low, whilst at high redshift the completeness is low and purity high. Only at a few intermediate redshifts are the completeness and purity simultaneously high, and this cross over is highly dependent on the adopted accretion threshold.

These trends can be explained by the average host halo mass of quasars and HzRGs. The massive halos that host HzRGs are the very peaks of the matter distribution at $z > 3.5$, tracing those regions that are most likely to form clusters (see Section 3.3.4.2), hence the high purity of the selection. By $z \sim 2$ halos of mass $\log_{10}(M \,/\, M_{\odot}) \sim 12.5$ are more numerous and do not necessarily coincide with protocluster regions, so the purity decreases, but the completeness rises sharply. We see no clear evidence for environmental triggering of HzRGs, as suggested by Hatch et al. (2014); instead, HzRGs occur within a narrow range of host halo masses, coincident with forming protocluster cores or groups (Chiang et al., 2017).

Similarly, at $z > 5$ the majority of high stellar mass (S_{MAS10}) galaxies reside in protoclusters (see Figure 3.2), so major mergers between such galaxies, triggers of quasar mode accretion, will predominantly occur in protocluster environments, hence the high purity of quasar tracers. This is true of both accretion cuts; the most luminous quasars at $z \sim 6$ do indeed reside in protoclusters, but there are far too few of them to trace an appreciable number of protoclusters. At later times there is also a population of massive galaxies in the field that may merge, reducing the purity. There are also less frequent mergers between massive galaxies in protoclusters once a dominant subhalo has formed at the core, which could be responsible for the plateau in completeness at low redshifts.

Orsi et al. (2016) find similar trends in their model; they observe that half of all HzRGs at $z = 2.2$ have cluster descendants, whereas in our model the fraction is approximately between a third and a half, depending on the accretion threshold. They also find 19% of quasars have cluster descendants, similar to our value of $\sim 21\%$ for the standard accretion threshold, but slightly lower than the conservative cut. Observationally, Venemans et al. (2007) find that 75% of powerful HzRGs in the redshift range $2 \leqslant z \leqslant 5$ reside in protoclusters, which agrees approximately with the mean AGN fraction in this range for both accretion thresholds. They use a $\sim 3 \times 3$ Mpc aperture, much smaller than R_{C}; the analysis in Section 3.3.5.2 suggests that measuring overdensity around HzRGs on this scale

will be biased lower, which makes their high measured protocluster fraction somewhat surprising, however they do adopt a more lenient protocluster definition (a factor of 2-5 overdense compared to the field; Figure 3.12 suggests an overdensity > 8 is required) and observe very powerful HzRGs which may be biased toward high mass protoclusters with higher probabilities. The Clusters Around Radio-Loud AGN (CARLA) survey (Wylezalek et al., 2013) found 66% of HzRGs reside in overdense regions at $z \sim 2.4$ (Hatch et al., 2014), approximately equal to the conservative accretion threshold, and Wylezalek et al. (2013) find 55% of HzRGs are overdense by 2σ, and 10% by $\geqslant 5\sigma$ (for $1.2 < z < 3.2$), which, if we assume that the lower overdensity limit corresponds to true protoclusters, matches our conservative accretion threshold, and the results of Orsi et al. (2016).

How effective are AGN as biased tracers of protoclusters? Our model suggests that it depends strongly on redshift. At high redshift, HzRGs act as reliable tracers of protocluster regions but will not reveal the presence of all protoclusters, whereas quasars reside in a more diverse range of environments. At lower redshifts almost all protoclusters have at least one AGN, but most AGN do not reside in protoclusters. At extremely high redshifts, Figure 3.2 suggests that using masssive galaxies as tracers will lead to the identification of a much more complete sample of protoclusters compared to using AGN, though it should be noted that such galaxies will typically exhibit observable AGN activity too. We leave the investigation of whether AGN-hosting protoclusters are a distinguishable population for future work.

3.4 Discussion

In Section 3.3.4.1 we presented an improved procedure for predicting the fate of observed galaxy overdensities. To demonstrate, we apply the technique to a number of observational candidates in the literature. Table 3.3 lists estimated protocluster probabilities and descendant masses for 13 protocluster candidates from the literature, each of which have been studied in Chiang et al. (2013). We also apply the technique to the first 12 candidates presented in the Candidate Cluster and Protocluster Catalogue (CCPC) compiled in Franck & McGaugh (2016a), shown in Table 3.4; this catalogue, whilst heterogenously selected, uses smaller, regular ($2R \sim D_C$) apertures to measure overdensity, and provides predictions for the protocluster probability and decendant mass derived from Chiang

Table 3.3: Estimated protocluster probabilities for candidates from the literature. All candidate estimates use the $S_{\rm SFR}$ selection, and combine the Proto and PartProto selections in the protocluster definition. Descendant mass estimates are omitted where protocluster probabilities are low.
Notes: (a) Redshift. (b) Full width redshift uncertainty. (c) Aperture length corresponding to redshift uncertainty. (d) Observation window area in square arc minutes. (e) Aperture radius giving equal area to the observation window. (f) Measured galaxy overdensity within the specified aperture. (g,h) Mean completeness and purity for each selection, and $5^{\rm th} - 95^{\rm th}$ percentile range. We use the lower percentile as our value for $C_{\rm lim}$ and $P_{\rm lim}$. (i) Derived protocluster probability. (j) Descendant masses estimated using our fitting procedure.
References: (1) Venemans et al. (2007) (2) Steidel et al. (2005) (3) Hatch et al. (2011b) (4) Tanaka et al. (2011) (5) Venemans et al. (2005) (6) Matsuda et al. (2005) (7) Steidel et al. (2000) (8) Yamada et al. (2012) (9) Venemans et al. (2002) (10) Venemans et al. (2004) (11) Ouchi et al. (2005) (12) Toshikawa et al. (2012)

Name	z [a]	Δz [b]	D [c] cMpc	Window [d] arcmin^2	R [e] cMpc	δ_g [f]	$C_{\rm lim}$ [g]	$P_{\rm lim}$ [h]	$P_C(\mathrm{S_{SFR1}})$ [i]	$\log_{10}(M_{z=0}/M_\odot)$ [j]
PKS 1138-262[1]	2.16	0.053	72.6	49	6.36	3^{+2}_{-2}	$0.92^{1.0}_{0.60}$	$0.28^{0.50}_{0.15}$	50%	14.530
HS1700-FLD[2]	2.3	0.03	38.7	64	7.52	$6.9^{+2.1}_{-2.1}$	$0.98^{1.0}_{0.72}$	$0.34^{0.59}_{0.18}$	100%	15.089
4C 10.48[3]	2.35	0.046	58.0	6.25	2.37	11^{+2}_{-2}	$0.3^{0.6}_{0.08}$	$0.56^{0.86}_{0.26}$	1.0%	-
4C 23.56[4]	2.48	0.035	41.8	28	5.16	$4.3^{+5.3}_{-2.6}$	$0.8^{0.97}_{0.44}$	$0.47^{0.72}_{0.26}$	55%	14.557
MRC 0052-241[1,5]	2.86	0.054	55.6	49	7.32	$2^{+0.5}_{-0.4}$	$0.94^{1.0}_{0.62}$	$0.34^{0.59}_{0.18}$	55%	14.497
MRC 0943-242[1,5]	2.92	0.056	56.4	49	7.39	$2.2^{+0.9}_{-0.7}$	$0.94^{1.0}_{0.63}$	$0.34^{0.58}_{0.18}$	55%	14.430
SSA22-FLD[6,7,8]	3.09	0.066	62.5	81	9.74	5^{+2}_{-2}	$1.0^{1.0}_{0.83}$	$0.21^{0.44}_{0.11}$	29%	-
MRC 0316-257[1,5]	3.13	0.049	45.8	49	7.62	$2.3^{+0.5}_{-0.4}$	$0.95^{1.0}_{0.65}$	$0.37^{0.62}_{0.20}$	59%	14.486
TN J2009-3040[1,5]	3.16	0.049	45.3	49	7.65	$0.7^{+0.8}_{-0.6}$	$0.95^{1.0}_{0.65}$	$0.37^{0.62}_{0.20}$	2.4%	-
TN J1338-1942[1,5,9]	4.11	0.049	33.5	49	8.52	$3.7^{+1.0}_{-0.8}$	$0.97^{1.0}_{0.70}$	$0.43^{0.70}_{0.23}$	71%	14.729
TN J0924-2201[10]	5.19	0.073	37.6	49	9.25	$1.5^{+1.6}_{-1.0}$	$0.98^{1.0}_{0.73}$	$0.40^{0.68}_{0.21}$	30%	-
SXDF-Object 'A'[11]	5.7	0.099	45.3	36	8.18	$3.3^{+0.9}_{-0.9}$	$0.94^{1.0}_{0.63}$	$0.44^{0.72}_{0.23}$	79%	14.651
SDF-12[3]	6.01	0.05	21.4	36	8.31	16^{+7}_{-7}	$0.95^{1.0}_{0.64}$	$0.62^{0.87}_{0.36}$	100%	> 15.3

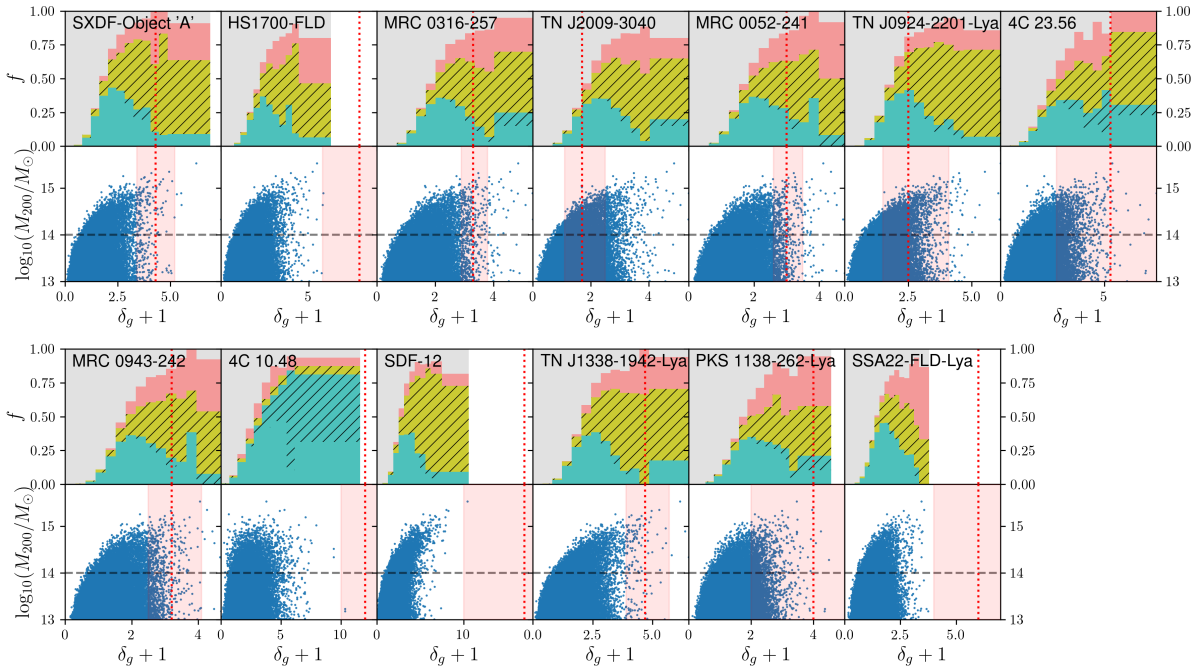

Figure 3.14: *Top panels:* Probability distributions for each candidate from Table 3.3 (labelled) for 100 000 random regions with the same dimensions as the given candidate. Probabilities are labelled identically to Figure 3.6. The observationally measured overdensity is shown as a vertical dotted red line; where the overdensity exceeds the maximum overdensity from the random sampling, we show white space. *Bottom panels:* Descendant mass against overdensity measured in the candidate aperture for all halos with $M/M_\odot > 10^{13}$. The cluster mass threshold is shown as the horizontal black dashed line. Uncertainties in the observationally measured overdensity are shaded in red.

Table 3.4: Estimated protocluster probabilities for the 12 strongest candidates from the CCPC catalogue (Franck & McGaugh, 2016a).
Notes: (a) Redshift. (b) Measured galaxy overdensity within a cylindrical aperture with radius $R = 10$cMpc, and depth $2\,\sigma_z = D$. (c) Full width redshift uncertainty. (d) Aperture length corresponding to redshift uncertainty. (e,f) Mean completeness and purity for each selection, and $5^{\rm th} - 95^{\rm th}$ percentile range. We use the lower percentile as our value for $C_{\rm lim}$ and $P_{\rm lim}$. (g) Protocluster probabilites from Franck & McGaugh (2016a), calculated using Figure 8 from Chiang et al. (2013) using the same selection (S_{S10}) (h) Derived protocluster probabilities, combining the Proto and PartProto selections. (i) Descendant masses estimated using our fitting procedure. (j) Coefficient of determination.
References: (1) Venemans et al. (2007) (2) Møller & Fynbo (2001) (3) Steidel et al. (1998) (4) Ellison et al. (2001)

Name	z [a]	δ_g [b]	σ_z [c]	D (cMpc) [d]	$C_{\rm lim}$ [e]	$P_{\rm lim}$ [f]	P_C (F&M) [g]	P_C(S_{S10}) [h]	$\log_{10}(M_{z=0}/M_\odot)$ [i]	R^2 [j]
CCPC-z27-002	2.772	11.02 ± 6.9	0.007	14.9	$1.0^{1.0}_{0.8}$	$0.89^{1.0}_{0.54}$	100%	75%	14.47	0.63
CCPC-z29-001	2.918	11.21 ± 4.76	0.005	10.08	$1.0^{1.0}_{0.67}$	$1.0^{1.0}_{0.64}$	100%	46%	14.28	0.63
CCPC-z29-002[1]	2.919	12.91 ± 4.55	0.009	18.12	$1.0^{1.0}_{0.82}$	$0.86^{1.0}_{0.5}$	100%	83%	14.67	0.61
CCPC-z30-001[2]	3.035	18.78 ± 10.14	0.005	9.64	$1.0^{1.0}_{0.67}$	$1.0^{1.0}_{0.67}$	100%	74%	14.61	0.61
CCPC-z30-003[3]	3.096	12.28 ± 2.42	0.008	15.10	$1.0^{1.0}_{0.8}$	$0.89^{1.0}_{0.55}$	100%	74%	14.55	0.63
CCPC-z31-003[1]	3.133	9.80 ± 2.77	0.008	14.92	$1.0^{1.0}_{0.8}$	$0.89^{1.0}_{0.55}$	100%	48%	14.39	0.63
CCPC-z31-004	3.146	7.59 ± 4.65	0.006	11.14	$1.0^{1.0}_{0.71}$	$1.0^{1.0}_{0.62}$	85%	14%	14.09	0.63
CCPC-z31-005[1]	3.152	17.77 ± 9.19	0.007	12.96	$1.0^{1.0}_{0.75}$	$0.92^{1.0}_{0.58}$	100%	86%	14.72	0.64
CCPC-z32-002	3.234	13.11 ± 8.63	0.003	5.40	$0.8^{1.0}_{0.3}$	$1.0^{1.0}_{0.67}$	100%	24%	14.11	0.49
CCPC-z33-002[4]	3.372	7.44 ± 4.47	0.008	13.74	$1.0^{1.0}_{0.78}$	$0.91^{1.0}_{0.57}$	85%	42%	14.17	0.63
CCPC-z35-001	3.597	10.18 ± 8.05	0.003	4.80	$0.6^{1.0}_{0.22}$	$1.0^{1.0}_{0.67}$	100%	1%	13.80	0.32
CCPC-z36-001	3.644	23.50 ± 14.39	0.003	4.72	$0.6^{1.0}_{0.2}$	$1.0^{1.0}_{0.67}$	100%	72%	14.12	0.31

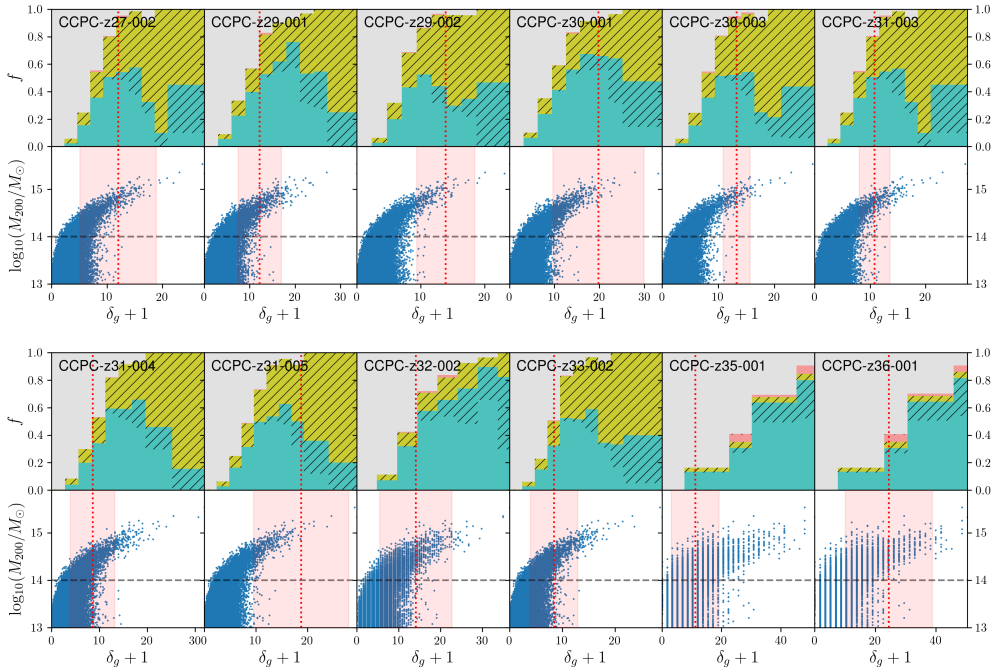

Figure 3.15: As for Figure 3.14, but for the first 12 candidates from the Candidate Cluster and Protocluster Catalogue (CCPC) (Franck & McGaugh, 2016a) listed in Table 3.4 and discussed in Section 3.4.

et al. (2013) that facilitate a direct comparison with our method. In both cases we use an aperture with the same dimensions as the observations.[12] For the candidates in Table 3.3 we use the S_{SFR1} selection, since all of these candidate overdensities are measured with star forming galaxies, whereas for Table 3.4 we use the S_{MAS10} selection identical to that used in Franck & McGaugh (2016a); they acknowledge that this selection does not correspond exactly with the selection used to identify their candidates, but represents a conservative lower estimate (if the selection does include lower mass galaxies this would boost the overdensity measurement, and therefore the corresponding probabilities) Each candidate is classified according to the 5^{th} percentile of the completeness and purity of the protocluster population.

Many of the candidates in Table 3.3 are measured with large apertures ($> (30\,\mathrm{cMpc})^3$), which has a significant effect on derived descendant properties. The bottom panels of Figure 3.14 show the relationship between overdensity and descendant mass for all halos with $M/M_\odot > 10^{13}$ in our model for the same aperture as each of these candidates; it is clear that for many it is very difficult to distinguish the protocluster population from the field in overdensity space. 4C10.48 is measured within a particularly pathological aperture ($R \ll D_C$) that leads to almost no distinction between the populations. This effect can also be seen in the probability distributions in the top panels of Figure 3.14. Above intermediate overdensities the Proto probability actually *decreases* relative to the PartProto probability; if a large aperture happens to capture parts of two protoclusters, the overdensity will be boosted by both overdensities but the probabilities will be affected by the low completeness of each protocluster.

The measured overdensity for 4C10.48 is much larger than that seen in randomly sampled regions or surrounding protoclusters, and we see similarly high overdensities for $\mathrm{HS1700-FLD}$, $\mathrm{SSA22-FLD-Ly}\alpha$ and $\mathrm{SDF-12}$. We attribute these high overdensities to two primary effects. First, each of these candidates is measured within a large aperture, which can be susceptible to aperture effects; our approach cannot distinguish the capture of two protoclusters within an aperture, or the chance alignment along a filamentary structure that is not destined to fall within the virial radius of the cluster at $z = 0$. Second, the selection criteria is not identical to that used for each candidate; a more conservative selection criteria could lead to a substantial boost in overdensity

[12] where rectangular apertures are used, we approximate with a cylinder of equal volume

measurement (Chiang et al., 2013). Chiang et al. (2013) note that TN J2009 − 3040 is most likely a large group or low mass protocluster, and we come to a similar conclusion; Figure 3.14 shows that, whilst a number of protoclusters have a similar overdensity, a large number of groups also exhibit similar overdensities, which is reflected in the protocluster probability.

Figure 3.15 shows the probability and descendant mass distributions for the CCPC candidate apertures, listed in Table 3.4 with probabilities and descendant mass estimates. These candidates are typically measured with smaller apertures, which leads to greater distinction between protoclusters and the field, and high protocluster probabilities for sufficiently high overdensities; the majority are confirmed as protoclusters with high confidence. CCPC-z32-002 is assigned a lower protocluster probability since it lies close to the overdensity threshold below which protoclusters are difficult to distinguish, and CCPC-z35-001 is ruled out with high confidence; whilst there are protoclusters with the same overdensity, the vast majority of objects with this overdensity have relatively low halo masses.

All of our results are simulation dependent, though we note that the pipeline is not, so it can be run again using catalogues from other simulations. We also include protocluster regions in our calculation of the average field overdensity, so the field overdensity is an overestimate. However, typical observable measures of field overdensity use the region in the foreground and background of the protocluster as a proxy for the 'field' (Franck & McGaugh, 2016a,b); since protoclusters have no sharp edge (see Figure 3.4), this approach may inadvertantly sample the protocluster overdensity tail, boosting the 'field' overdensity. It's unclear to what degree these two effects cancel out.

3.5 Summary

We have used L-Galaxies to investigate the characteristics of galaxy protoclusters. Our findings are as follows:

- The completeness and purity of the protocluster galaxy population are maximised ($> 85\,\%$) at a radius of $R_{\rm C} \approx 10 \pm 2\,{\rm cMpc}$. This scale is insensitive to redshift and galaxy selections. Galaxy overdensities measured on $R_{\rm C}$ provide high discrimination between protoclusters and the field, particularly at high redshift, and overdensities

surrounding quasars and HzRGs are also best measured at $R_{\rm C}$ since AGN are not centrally located within protoclusters.

- Protocluster galaxies exhibit aspherical, prolate distributions, though this has little effect on their completeness and purity as measured within $R_{\rm C}$ due to the lower density of galaxies in the field on their outskirts. Redshift space distortions slightly boost the measured overdensity, since protocluster galaxies tend to be infalling due to the Kaiser effect.

- Using AGN as tracers at $z \gtrsim 5$ is accurate but highly incomplete. The most luminous quasars at $z \sim 6$ are correlated with protocluster regions, but there are too few of them to act as tracers.

- The most massive galaxies at all epochs preferentially appear in protocluster environments, and we see indirect evidence for the emergence of a red sequence in protoclusters through their greater asphericity and steeper completeness curves at $z \leqslant 3$.

- We have demonstrated a procedure for generating protocluster probabilities based on their measured galaxy overdensity that can be applied to irregular apertures. We apply it to a range of redshifts and selection criteria, and provide fits between overdensity and descendant cluster mass. Low mass protoclusters cannot be discriminated due to overlap in overdensity space with field regions.

We make all of the code used in this paper public, at https://github.com/christopherlovell/goa. It can be used to run the pipeline outlined in Section 3.3.4; we hope it will be of use to observers wishing to identify and characterise high-z galaxy overdensities.

3.6 Appendix

3.6.1 Overdensity Statistics

Figure 3.16 shows the effect of adjusting our free parameters, $C_{\rm lim}$ and $P_{\rm lim}$, whilst keeping a fixed aperture volume ($R = D/2 = 10\,$cMpc). Changing $C_{\rm lim}$ principally affects the

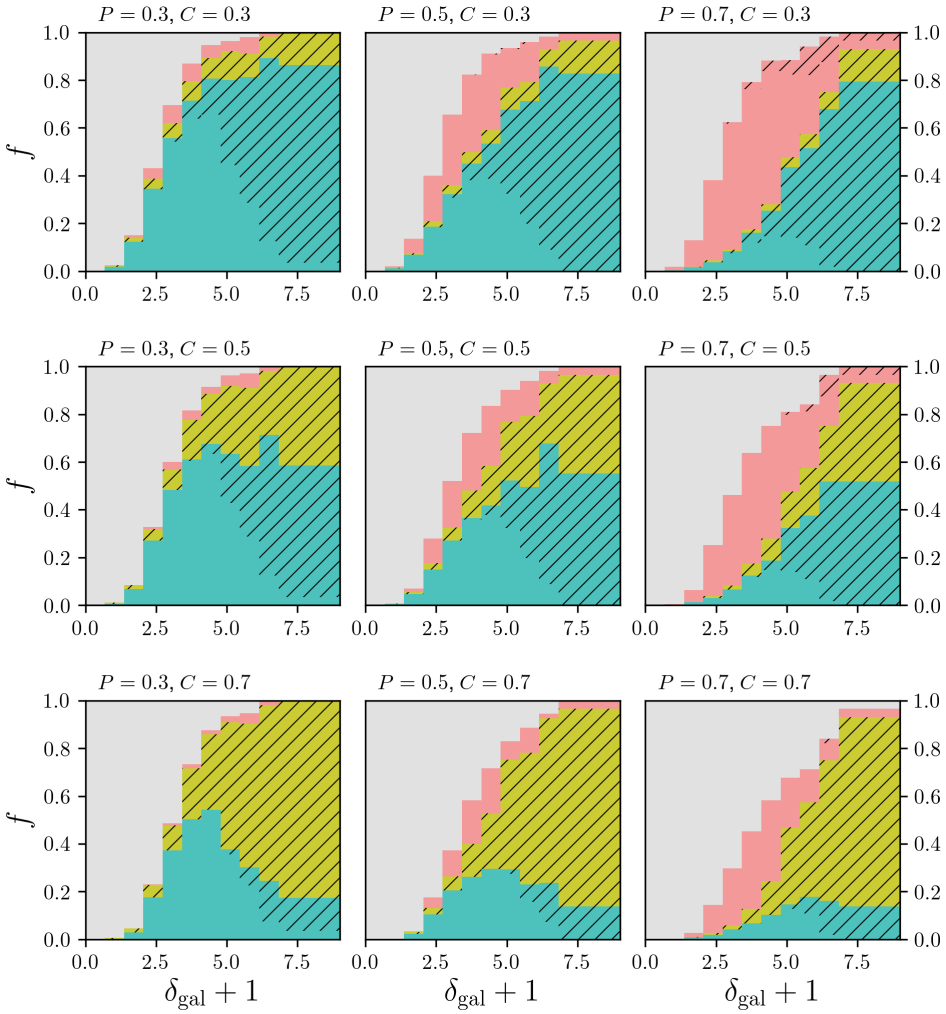

Figure 3.16: Fractional probability distributions for different choices of C_{lim} and P_{lim} (See Figure 3.6 for legend). In general, the higher the purity constraint, the more regions are classified as ProtoField, and the higher the completeness constraint, the more regions are classified as PartProto. Higher P_{lim} can also lead to higher Field probabilities.

ratio of probability of PartProto to Proto, and $P_{\rm lim}$ lowers the Proto probability for a given overdensity, and increases the ProtoField probability. A liberal choice of both $P_{\rm lim}$ and $C_{\rm lim}$ leads to high protocluster probabilities, but the probability of probing a field region at low overdensity is still high. Choosing both $P_{\rm lim}$ and $C_{\rm lim}$ conservatively leads to PartProto probabilities dominating. We recommend choosing values of $P_{\rm lim}$ and $C_{\rm lim}$ motivated by the completeness and purity of the protocluster population, given the aperture choice and selection.

4 Galaxy Protoclusters in the Cluster-Eagle project: evolution of the star-forming sequence

Christopher C. Lovell,[1] Peter A. Thomas,[1] David J. Barnes,[2] Yannick M. Bahé,[3] Stephen M. Wilkins[1] Scott T. Kay,[4]

4.1 Introduction

In this chapter we present a study of the star-forming sequence in galaxy protoclusters with the C-Eagle simulations.

Galaxy protoclusters, the high-redshift progenitors of galaxy clusters, contain some of the most massive, highly star-forming galaxies at $z > 2$. However, it is unclear whether, at a fixed stellar mass, protocluster galaxies have differing star formation rates than those in the field, and how this evolves with redshift. We present an investigation of the stellar mass-star formation rate relation, or star-forming sequence, in protoclusters simulated with full hydrodynamics. We utilise periodic box simulations from the Eagle project, and zoom simulations from the Cluster-Eagle project. Cluster-Eagle is comprised of 30 clusters with a range of $z = 0$ descendant masses, providing a large number of galaxies in both field and protocluster environments. Together these simulations allow us to study the dependence of the SFS on protocluster environment.

We also compare to a number of observational studies in both field and protocluster environments. Protocluster observations are notoriously difficult due to the difficulty of determining protocluster membership, particularly for quiescent objects. We therefore make tentative comparisons to a number of observed protocluster relations for the SFS, and

[1]Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton, BN1 9QH, UK
[2]Department of Physics, Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
[3]Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands
[4]Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK

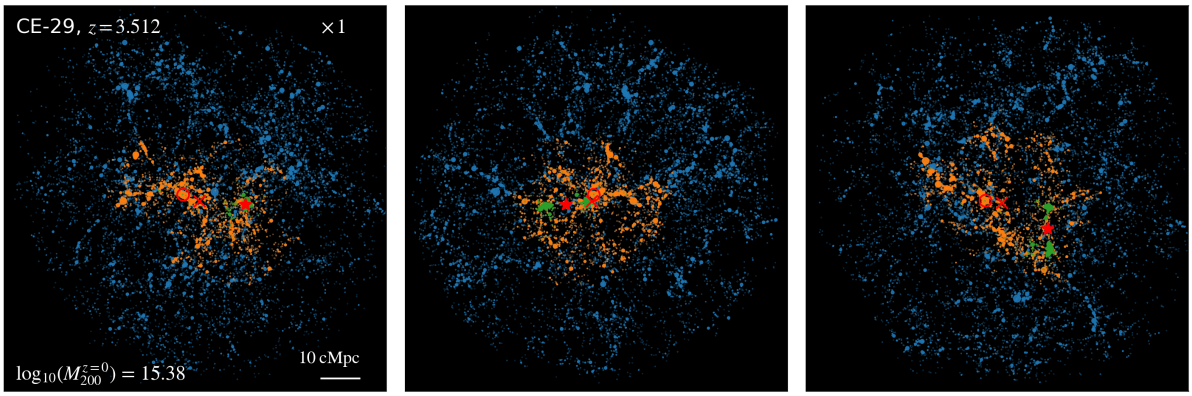

Figure 4.1: Galaxy distribution in a high descendant mass ($> 10^{15} M_{\odot}$) protocluster at $z = 3.5$ from three orthogonal perspectives. Points are scaled by the galaxy stellar mass. Protocluster galaxies are shown in orange, surrounding field galaxies in blue, and proto-BCG galaxies in green. The red circle indicates the most massive galaxy in the protocluster, the red cross the protocluster centre of mass, and the red star the proto-BCG centre of mass.

indicate where possible the systematic uncertainties and their effect on our conclusions.

The study is arranged as follows. In Section 4.2 we describe the C-EAGLE simulations and relevant definitions. We then discuss the main sequence in protoclusters (Section 4.3), comparing with detailed observational studies of well known protoclusters in Section 4.3.5, and investigate the scatter around the SFS in Section 4.3.6. In Section 4.4 we study the passive fraction, and compare to well studied protoclusters. Finally, in Section 4.5 we discuss our results, and state our conclusions in Section 4.6.

4.2 Methods

4.2.1 The Simulations

The Cluster-Eagle project applies the EAGLE model, described in detail in (Schaye et al., 2014; Crain et al., 2015), to cluster environments using the 'zoom' re-simulation technique (Katz & White, 1993; Tormen et al., 1997). Clusters, defined as objects with halo mass $\mathrm{M_{200}} / \mathrm{M_{\odot}} > 10^{14}$, are selected at $z = 0$ from the parent volume described in Barnes et al. (2017a), a $(3.2\ \mathrm{Gpc})^3$ dark-matter-only simulation using the Planck Collaboration et al. (2014) cosmology. It uses an identical resolution to the fiducial EAGLE simulation, with gas particle mass $m_g = 1.8 \times 10^6\ \mathrm{M_{\odot}}$, and a physical softening length of 0.7 kpc. We limit our analysis to galaxies sampled by at least 100 star particles, *e.g.* $\log_{10}(\mathrm{M_*} / \mathrm{M_{\odot}}) \geqslant 8.25$.

C-Eagle uses the AGNdT9 calibration of the Eagle model, which, compared to the Reference (Ref) model, uses a higher value for $C_{\rm visc}$, which controls the sensitivity of the BH accretion rate to the angular momentum of the gas, and a higher gas temperature increase from AGN feedback, ΔT. A larger ΔT leads to fewer, more energetic feedback events, whereas a lower ΔT leads to more continual heating.

Furlong et al. (2015) showed that high redshift galaxy properties in Eagle, such as the stellar density and galaxy stellar mass functions, are within observational bounds up to at least $z \sim 6$. The slope of the specific star formation rate relation matches observations, but the normalisation is lower by 0.5 dex at $z = 2$. This offset, first identified by Daddi et al. (2007), is present in other hydrodynamic simulations; we discuss the implications and possible resolutions in further detail in Section 4.5. The stellar mass content of $z = 0$ clusters in the C-Eagle simulations is in line with observations, though there is an offset of +0.3 dex in the stellar mass of the brightest cluster galaxy by $z = 0$ (Bahé et al., 2017). We dicuss the implications of this offset in the context of the results presented here in Section 4.5.

4.2.2 Definitions

Merger trees in C-Eagle are constructed using the Spiderweb tracing algorithm, described in detail in the Appendix of Bahé et al. (2019), which consistently tracks the baryonic component of galaxies through disruption and stripping events. We use these to make the following definitions for different galaxy populations throughout the rest of the paper:

- **Protocluster galaxies** are defined as the progenitors of all galaxies that lie within $R_{200}^{\rm crit}$ of a cluster at $z = 0$ in the C-Eagle simulations.
- **Proto-BCG galaxies** are defined as all progenitors of the central galaxy in the cluster, ignoring satellites.

Field galaxies are strictly defined as all galaxies that will **not** end up within the virial radius of the cluster at $z = 0$. However, protoclusters at high redshift are aspherical, with a prolate distribution, and there is often mixing of field and protocluster galaxies at the edge (Lovell et al., 2018). In order to select a cleaner sample of field galaxies we therefore define them as follows:

- **C-Eagle field galaxies** are defined as any galaxies that lie outside a bounding sphere centred on the median of the protocluster galaxy coordinates, with radius equal to the maximum protocluster galaxy distance, plus 1 cMpc.

Bahé et al. (2017) found that galaxies in the outskirts of clusters had elevated stellar-mass fractions and dark-matter concentrations compared to the field. We therefore also use the 50 Mpc periodic simulation using the AGNdT9 physics as an unbiased field comparison region. This volume contains a single cluster mass halo at $z = 0$, which we remove following the same bounding-sphere condition as above.

- **Periodic field galaxies** are defined as all galaxies in the 50 Mpc periodic simulation that lie outside the bounding-sphere of the single protocluster present in the volume.

The periodic simulations have output snapshots at slightly different times to the C-Eagle simulations; we match to the nearest snapshot where available.

The 100 Mpc Reference simulation has a larger volume, but uses different model parameters. The spatial distribution of galaxies in a high descendant-mass protocluster is shown in Figure 4.1.

4.3 The Star-Forming Sequence

We now use our galaxy samples to study the dependence of the SFS on protocluster environment. To remove quiescent galaxies we impose a specific-star formation rate (sSFR) cut that excludes those galaxies whose current star formation is insufficient to double the mass of the galaxy within twice the current age of the universe,

$$\mathrm{sSFR} > \frac{1}{2 \times t_{\mathrm{age}}} \ ,$$

which leads to an evolving threshold for quiescence with redshift, shown in Figure 4.2. We tested using different thresholds (mass multiples of $\times\frac{3}{2}$ and $\times 3$) and found that all our results are insensitive to the multiple of mass chosen except for the passive fraction, discussed in greater detail in Section 4.4. Observations typically use UVJ colour to discriminate quiescent objects (e.g. Whitaker et al., 2011); at $z \sim 2$, this leads to a similar threshold for quiescence as a sSFR cut (Fang et al., 2018).

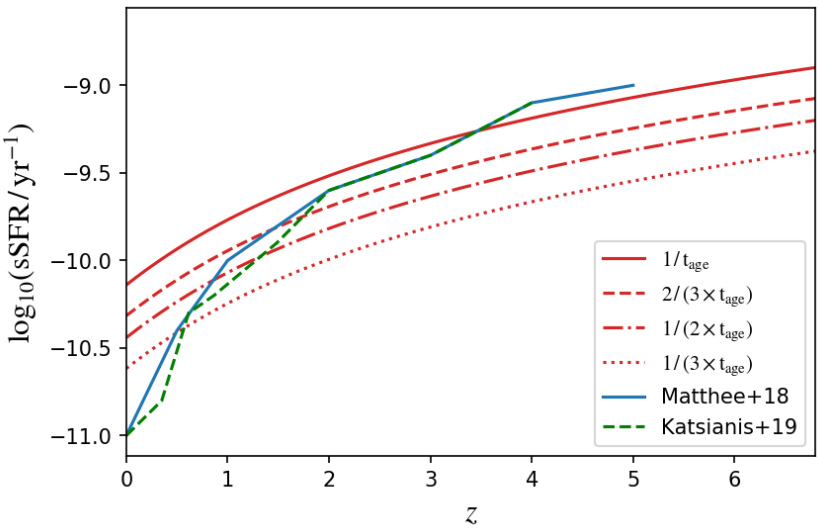

Figure 4.2: Our evolving sSFR threshold for quiescence. Shown is the threshold for different fractions of the current age of the universe. We use the mass doubling time throughout the rest of the paper, but note that our results are insensitive to the chosen ratio. The thresholds used in Matthee et al. (2017) & Katsianis et al. (2019) are shown for comparison.

Figure 4.3 shows the SFS from $z = 1.5$ up to $z = 7$, for protoclusters and both field regions. The normalisation is higher at $z \sim 7$, and decreases over cosmic time. Both field region selections show similar behaviour at all redshifts, but the periodic simulation does not extend to high stellar masses due to the smaller box size (lower by an order of magnitude). It is apparent that there is some similarity between the protocluster and field SFS; we discuss this in further detail throughout the rest of this section.

Figure 4.3 also shows comparisons with recent observational relations at high redshift. Each of these observations does not represent a dedicated field or protocluster environment, so should not be compared with either simulated environment explicitly; they should instead be interpreted as an indication of the combined SFS behaviour across all environments. We see an offset in the normalisation between the simulations and the observations at $z \sim 2$ of $\sim +0.3$ dex, first noted for the periodic simulations by Furlong et al. (2015). This discrepancy is not unique to EAGLE (Davé, 2008; Sparre et al., 2015; Donnari et al., 2019), and is remarkably consist across different simulations, both semi-analytic and hydrodynamic, employing very different subgrid physics recipes (Katsianis et al., 2016). We discuss this offset and its implications in further detail in Section 4.5.

Converted to a Chabrier IMF according to the offsets taken from Zahid et al. (2012).

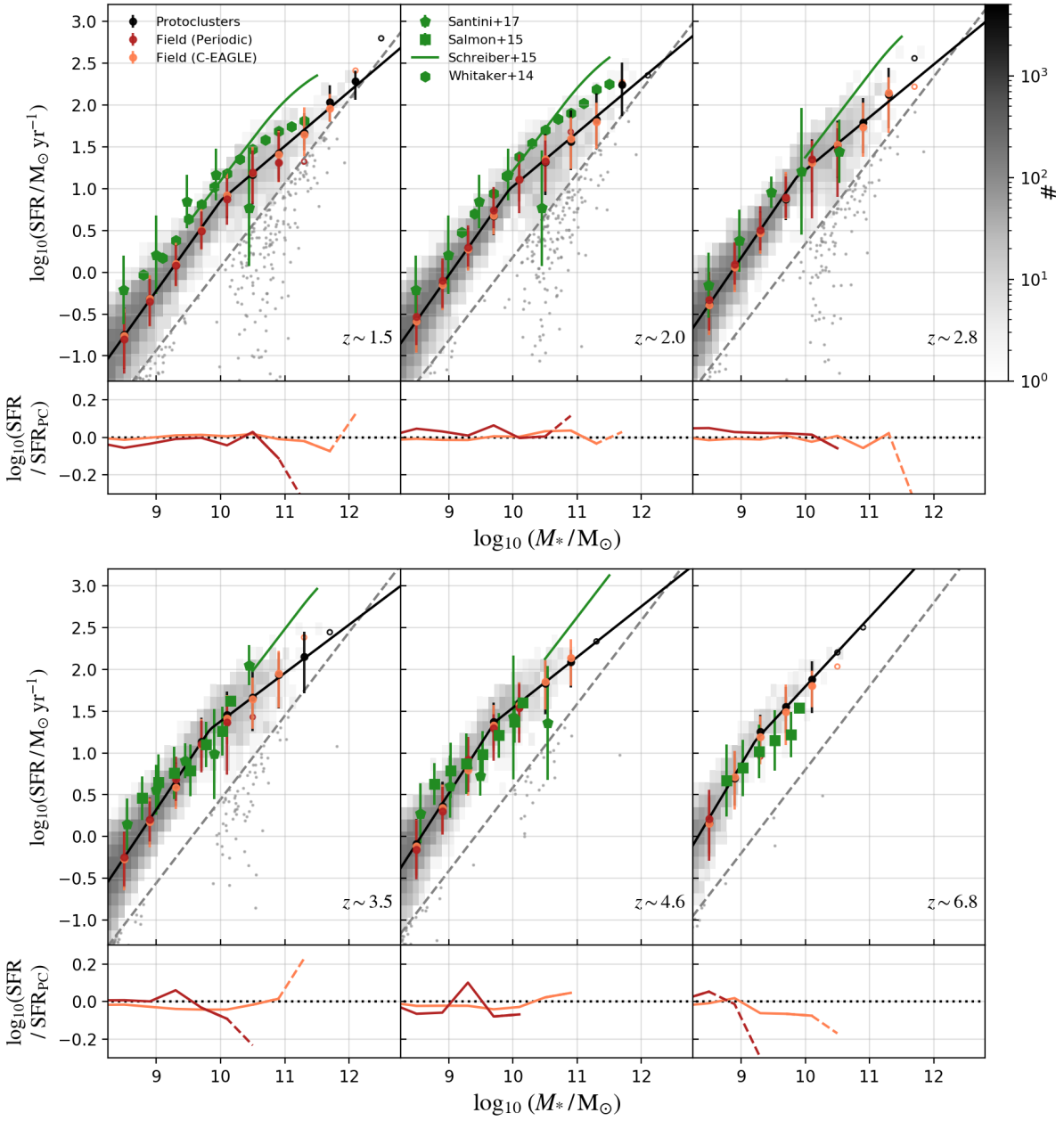

Figure 4.3: *Upper panels:* the star-forming sequence (SFS) for centrals over the redshift range $1.5 \leqslant z \leqslant 7$. The grey 2D histogram shows the distribution of protocluster galaxies on the SFS; the sSFR cut is shown as the grey dashed line, and individual quiescent galaxies below it as the grey scattered points. The black line shows the piecewise-fit to the protocluster relation, and black points show binned medians, with error bars giving the 10th-90th percentile spread. The yellow and gold points show the median relation for the field taken from the periodic AGNdT9 simulation, and the outskirts of the C-EAGLE re-simulations, respectively. Bins containing fewer than 10 objects are shown with non-filled points. Observational relations from Whitaker et al. (2014) (diamonds), Schreiber et al. (2015) (solid line), Salmon et al. (2015) (squares) and Santini et al. (2017) (pentagons) are shown in green. *Lower panels:* The log-ratio of the median relation in each field population to the protocluster relation. Bins containing fewer than 10 galaxies are shown with dashed lines.

At $z > 3$ the normalisation at the turnover mass is in good agreement with the observational constraints, however the Schreiber et al. (2015) results are in tension at the high mass end, with both the normalisation and the slope. The Whitaker et al. (2014) results also have a shallower low mass slope at these redshifts. There are currently no mass-complete observational constraints of the SFS at $z > 6.5$; we choose to show the $z = 6$ results from Salmon et al. (2015), which are in good agreement, but with a slightly lower normalisation which may be due to the redshift offset. This situation will change in the coming years with the launch of the James Webb Space Telescope, which will provide unprecedented infrared sensitivity to allow the characterisation of stellar masses and SFRs for the large number of protocluster candidates discovered recently at these redshifts (Higuchi et al., 2018).

4.3.1 Fit to the star-forming sequence

To investigate in detail the apparent similarity on the SFS between protocluster and field galaxies we first fit their distributions. Both the field and protocluster main sequence show evidence for a turnover at high masses ($\sim> 10^{9.5}\mathrm{M}_{\odot}$). To account for this we fit a two-part piecewise linear relation to the distribution, for stellar mass re-normalised at $10^{9.7}\mathrm{M}_{\odot}$,

$$\log_{10}(\mathrm{SFR}) = \alpha_1 \log_{10}(M_* \,/\, 10^{9.7}\mathrm{M}_{\odot}) + \beta_1 \quad M_* \leqslant M_{*,0} \tag{4.1}$$

$$\log_{10}(\mathrm{SFR}) = \alpha_2 \log_{10}(M_* \,/\, 10^{9.7}\mathrm{M}_{\odot}) + \beta_2 \quad M_* \geq M_{*,0}\ , \tag{4.2}$$

where α_1 is the low-mass slope, α_2 is the high-mass slope, and $M_{*,0}$ is the turnover mass in log-solar masses. The normalisation at the turnover, β_0, is then given by

$$\beta_0 = \beta_2 + \alpha_2\, M_{*,0} \tag{4.3}$$

$$= \beta_1 + \alpha_1\, M_{*,0}\ . \tag{4.4}$$

We use the SCIPY implementation of non-linear least squares to perform the fit, combined with a non-parametric bootstrap approach for estimating parameter uncertainties. The bootstrap is implemented as follows: we select, with replacement, 500 times from the

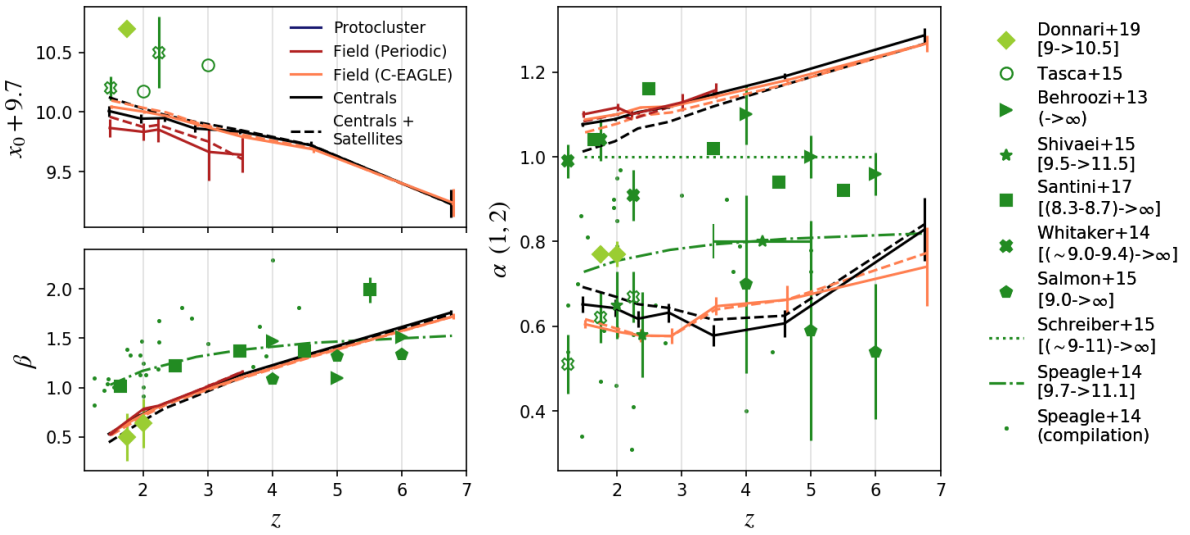

Figure 4.4: Parameters of the star-forming sequence (SFS) fit for protoclusters (black), periodic field regions (red), C-EAGLE field regions (orange), split into centrals only (solid) and including satellites (dashed). Errors on the parameters are derived from a non-parametric bootstrap analysis, computed as the 1σ spread in the bootstrap distributions. *Left, top:* the SFS turnover mass, $x_0 + 9.7$, in log solar masses. *Left, bottom:* the SFS normalisation, β_0. *Right:* the SFS low- and high-mass slopes, α_1 and α_2, respectively. We show these together for easier comparison with the range of observational constraints (dark green); where a piecewise relation has been used instead of a single linear relation in the observations, the high mass slope and turnover are shown with non-filled markers. Results from other recent simulations are shown in light green. Further details on each individual study are provided in Section 4.3.1.

original data, each resample being the same size as the original data. We then fit each sample independently; parameter estimates are given by the median of the resampled fit distributions, and uncertainties are given as the 1σ spread in the distributions (unless otherwise stated). The fit for protoclusters is shown in Figure 4.3 as the solid black line; the two part relation does a good job of fitting the low- and high-mass behaviour at all redshifts.

Figure 4.4 shows the evolution of the following fit parameters: turnover mass ($M_{*,0} + 9.7$), normalisation (β), low mass slope (α_1) and high mass slope (α_2). Where there are too few galaxies in a simulation at a given redshift to provide a good fit we omit the result (*e.g.* the periodic AGNdT9 box at $z > 3.5$). Both field and protocluster environments show similar evolution in their fit parameters, however there are notable differences, particularly when including satellite galaxies. For example, in both environments the normalisation and low-mass slope fall with decreasing redshift. However, when including satellites, both parameters are significantly (outside of the bootstrap errors) lower in protoclusters at

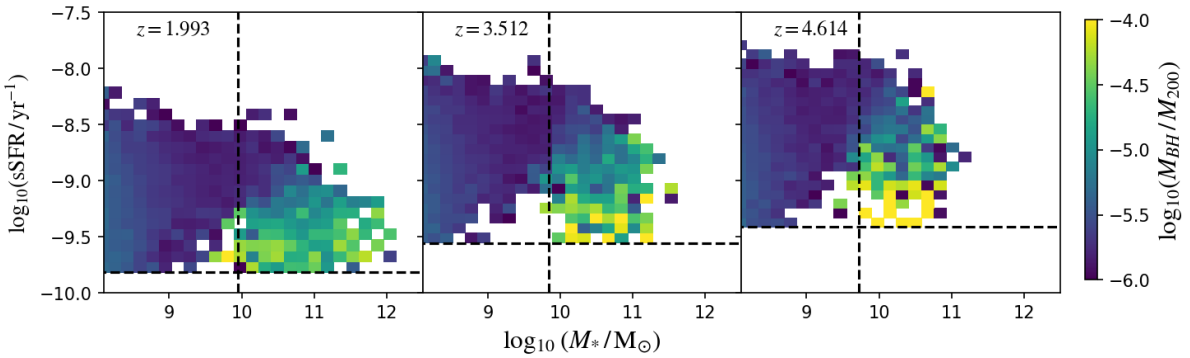

Figure 4.5: The specific star-formation rate-stellar mass relation for protocluster galaxies shown as a 2D histogram. All bins populated with at least a single object are shown. The colour shows the mean in that bin of the ratio of the black hole mass to the halo mass. The horizontal dashed line shows the sSFR cut for quiescence. The vertical dashed line shows the turnover mass for the protocluster star-forming sequence. Above the turnover mass there is a clear gradient in black-hole to halo mass ratio, at fixed stellar mass.

$z < 3$ compared to the field. We interpret this as additional environmental quenching of satellite galaxies in protocluster environments.

The turnover mass increases with decreasing redshift in both environments. There is a slight negative offset in the turnover mass measured in the periodic field region compared to C-Eagle, however this is within the bootstrap errors. It has been suggested that the turnover in the periodic Eagle volumes at lower redshift ($z < 2$) is due to the onset of AGN feedback (Matthee et al., 2017). To investigate whether this is also causing the turnover in the protocluster environment, Figure 4.5 shows the distribution of central protocluster galaxies on the stellar mass - specific star formation rate plane, coloured by central black hole mass. There is a clear gradient with black hole mass above the turnover mass, which suggests a similar origin for the turnover behaviour. The redshift evolution in the turnover can then be understood in terms of the evolving mass at which AGN feedback becomes dominant. (Matthee & Schaye, 2019) propose that the evolving stellar-halo mass relation is responsible for the evolving mass at which AGN 'switch on', as galaxies at fixed stellar mass reside in more massive halos at $z = 2$ compared to $z = 0$.

In both environments we see a fall in the high-mass slope, from $\alpha_2 \sim 0.8$ at $z > 6$, to $\alpha_2 \sim 0.65$ at $z < 2$. At $z > 4$ both environments high-mass slopes are within the bootstrap errors, however below this redshift they are significantly discrepant, with the protocluster high-mass slope tending to be greater than that in the field. This suggests that, given the near-identical normalisation at the turnover mass, high-mass galaxies in protoclusters

have *higher SFR* than those in the field.

We also show a number of observed fits to the SFS (Behroozi et al., 2013b; Whitaker et al., 2014; Tasca et al., 2015; Shivaei et al., 2015; Salmon et al., 2015; Schreiber et al., 2015; Santini et al., 2017), including the compilation of pre-2014 measurements from Speagle et al. (2014). As in Figure 4.3, these do not represent dedicated field or protocluster observations, but a combination of environments. They are a combination of single- and double-power law measurements; the former we show as filled points, the latter we show the low mass slope as filled points and the high mass slope as non-filled. For all observations we quote the approximate lower mass completeness limit for the whole fit in the legend. We do not quote the mass range of each measurement in the Speagle et al. (2014) results, but show the slope and normalisation of each study to give a qualitative indication of the intra-study distribution.

The observed slope shows considerable scatter spanning the range $\sim 0.4 - 1.2$. The majority of these relations quote a single power law relation, except Whitaker et al. (2014) (crosses) who quote a piecewise relation which shows reasonably good agreement with our low- and high-mass slopes up to $z \sim 2.5$. For the single power law relations the slope seems to be correlated with the lower mass limit of the survey. Where the lower mass completeness limit evolves with redshift this manifests as a decrease in the measured slope with redshift (e.g. Santini et al., 2017; Behroozi et al., 2013b; Salmon et al., 2015). Where the mass completeness lower limit is high (e.g. Shivaei et al., 2015) the slope tends to be shallower, in line with our measured high mass-slope.

This is illustrated more explicitly in Figure 4.6, which shows a negative correlation between the estimated lower-mass limit of the given observational survey and the value of the measured slope for a single power law. This suggests that many high redshift surveys, where the mass completeness does not extend to very low masses, are only probing the SFS at stellar masses above the turnover, and the measured slopes do not represent a universal relation for all masses.

For the turnover mass there are scarce observational constraints at $z > 3$, but a few studies have found and constrained the turnover at $z \sim 2$ (Tasca et al., 2015; Whitaker et al.,

For the (Whitaker et al., 2014) piecewise relation we plot the turnover mass as the lower mass limit for the high-mass power law

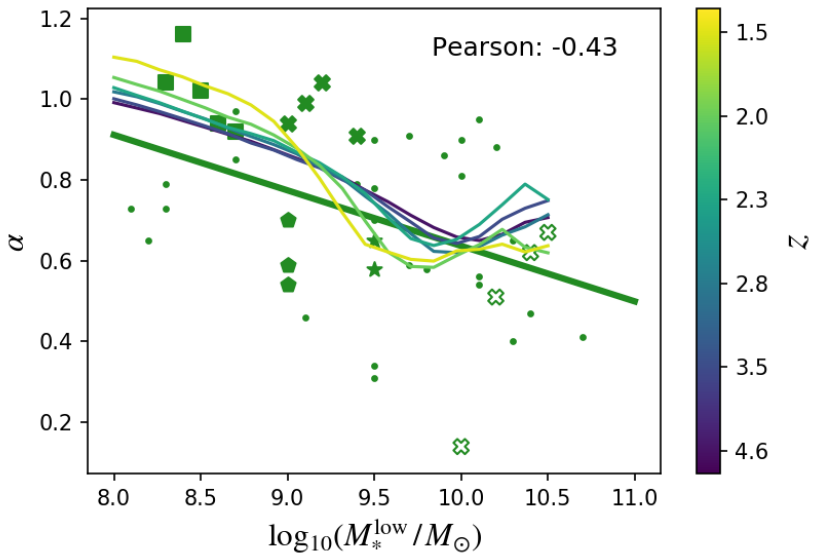

Figure 4.6: Estimated low-mass cut off against slope α for the observations plotted in Figure 4.4 (green). A linear fit to all the observations combined is shown with the solid line, and has a significant negative correlation (-0.43). The measured relation in the simulated protocluster sample for a linear fit with varying low mass slope is shown, coloured by redshift, and shows a similar relation.

2014). These studies suggest that the turnover mass *increases* with increasing redshift. This positive correlation has been explained as due to the downsizing paradigm, where more massive galaxies form their stars earlier (Neistein et al., 2006). We, suprisingly, see the *opposite* trend with redshift. As we have already discussed, the turnover evolution in EAGLE appears to be linked to the evolving stellar - halo mass relation. Another possible explanation is that, as galaxies become larger and less compact over time, AGN feedback is less efficient at curtailing star formation in low mass galaxies. It is unclear what observational effects could lead to the evolution seen in Tasca et al. (2015) & Whitaker et al. (2014).

The normalisation β shows reasonably good agreement with the observations at $z \geqslant 4$, but is offset around cosmic noon, as discussed at the start of this section. Figure 4.4 also shows results from Illustris-TNG; Donnari et al. (2019) fit a single power law, and find a slope between our high- and low-mass measurements, as expected since their mass completeness limit straddles the turnover mass. The normalisation is in very good agreement with our results, showing a similar tension with observations at cosmic noon.

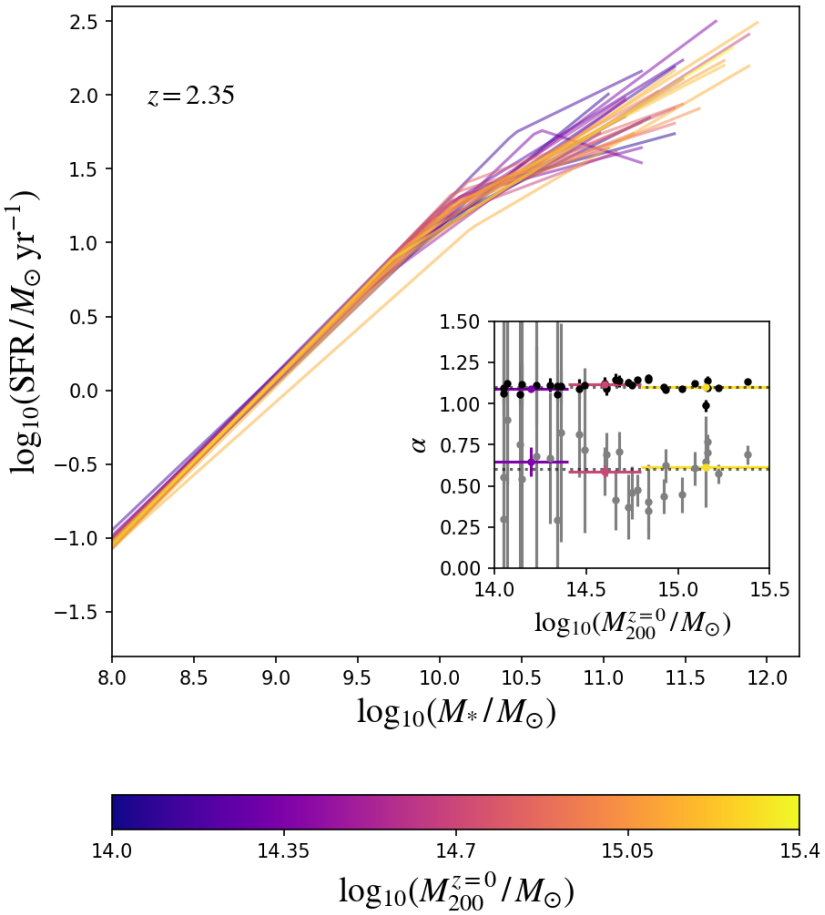

Figure 4.7: The star-forming sequence fit at $z = 2.35$ for each protocluster individually, coloured by descendant mass virial mass ($M_{200}^{z=0}$). *Inset:* The high- (grey) and low-mass (black) slope against descendant mass, with bootstrap 1σ errors. At low protocluster masses the uncertainties on the high-mass slope are large; mass-binned fits are shown in colour and show the mass-dependent trends clearer. There is no clear dependence of the fit on descendant mass.

4.3.2 The star-forming sequence of individual protoclusters

So far we have studied the galaxy population in all protoclusters combined. We now look at the differences in the SFS between individual protoclusters, with a range of descendant cluster masses. Figure 4.7 shows the piecewise fits to the SFS in each of our 30 protoclusters at $z = 2.35$, coloured by descendant virial mass. Below the turnover mass the behaviour of the SFS is very similar, but above this there is considerable variety in the high-mass slope between different protoclusters. The inset of Figure 4.7 shows the high- and low-mass slopes against descendant mass. The low-mass slopes show no dependence on descendant mass. It is difficult to see any trend in the high-mass slopes, due to the much larger errors at low descendant masses from the lack of high-mass galaxies. To better show the trends with mass we also show fits to protoclusters in three bins of descendant mass, ($\log_{10}(M_{200}^{z=0} \,/\, M_{\odot})$ = [14-14.4],[14.4-14.8],[14.8-]). The binned fits show no significant dependence of the high-mass slope on descendant mass.

What is striking in Figure 4.7 is the large scatter (~ 0.4 dex) in normalisation at the turnover mass. This suggests that protoclusters have varying evolutionary states at $z = 2.35$ regardless of their $z = 0$ descendant mass. We will see in Section 4.3.5 that this is partly due to the presence of dense groups within each protocluster. In particular, halo CE-27, with a descendant mass of $\log_{10}(\mathrm{M}_{200} \,/\, \mathrm{M}_{\odot}) = 15.15$, has a normalisation below the turnover 0.2 dex lower than any other protocluster; we will discuss this particular object, and its uniquely early-evolved galaxy population, in greater detail in future work.

4.3.3 Group-intergroup decomposition

Our analysis up to this point has shown that, whilst the protocluster and field SFS are similar in their general form and evolution, there are significant differences in their fits, particularly around cosmic noon ($1.5 < z < 3$) We also find a diversity of high-mass SFS behaviour for individual protoclusters that is not dependent on the descendant cluster mass. Protoclusters are web-like distributions at high redshift, composed of dense groups connected by filaments; it is therefore of interest to evaluate how the SFS depends on the presence and maturity of dense groups, and whether these denser environments promote or inhibit star formation at fixed stellar mass.

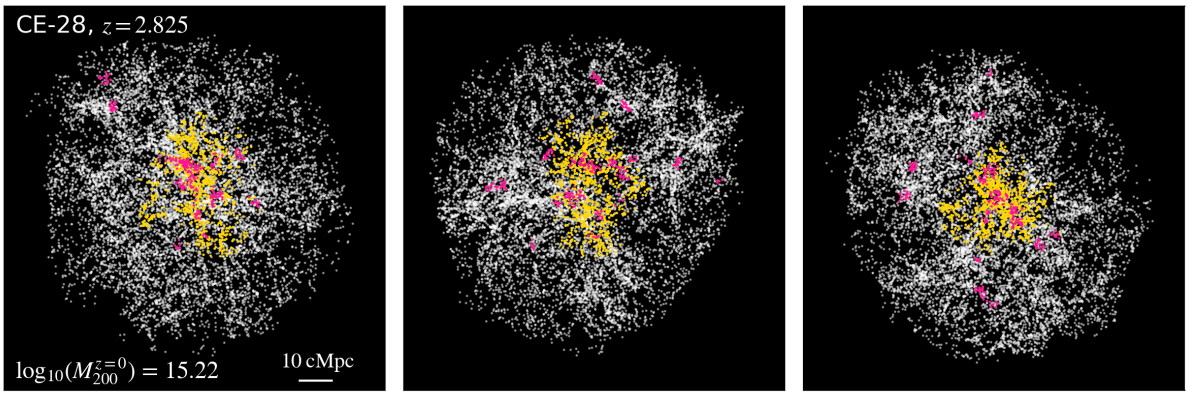

Figure 4.8: Galaxy distribution in a high descendant mass ($> 10^{15} M_{\odot}$) protocluster at $z = 2.8$ from three orthogonal perspectives, showing all galaxies (white), protoclusters galaxies (orange) and group galaxies (pink).

We use the full 3D information to measure the N^{th} nearest neighbour overdensity, characterised by the distance to the N^{th} nearest neighbour, r_N. This has been shown in numerical simulations to be reasonably correlated with the 2D surface overdensity for $N \geqslant 10$ (Shattow et al., 2013), and works sufficiently well on intra-halo scales (Muldrew et al., 2012), though aperture based approaches are better on large scales. We use $N = 30$ with a limiting scale of $r_N < 1.5$ cMpc, which corresponds to the 5^{th} percentile of r_N for the protocluster galaxy population at $z = 2.3$. Obviously, our group selection is sensitive to the chosen value of N and r_N; we chose a length that extended beyond the virial radius of the most massive halos at these redshifts to ensure we were not just identifying collapsed structures, but dense agglomerations of multiple halos.

Figure 4.8 shows the galaxy distribution in a high descendant-mass protocluster at $z = 2.8$ with the group galaxies highlighted. The algorithm identifies groups outside of the protocluster, which may collapse to form group-mass objects at $z = 0$, however we ignore these objects for now. Not all protoclusters contain groups at all redshifts, and the group fraction is lower in lower descendant mass protoclusters. The top panel of Figure 4.10 shows the fraction of group galaxies in the total protocluster galaxy population, as a function of redshift. Since we use a fixed group definition the group fraction increases with redshift, due to the increasing density of halos in the collapsing cluster.

Figure 4.9 shows the SFS at $z = 2.8$ for all protoclusters decomposed into group and intergroup populations, where a group has been identified. The points show the binned

We use a nearest neighbour approach for consistency with Shimakawa et al. (2017b)

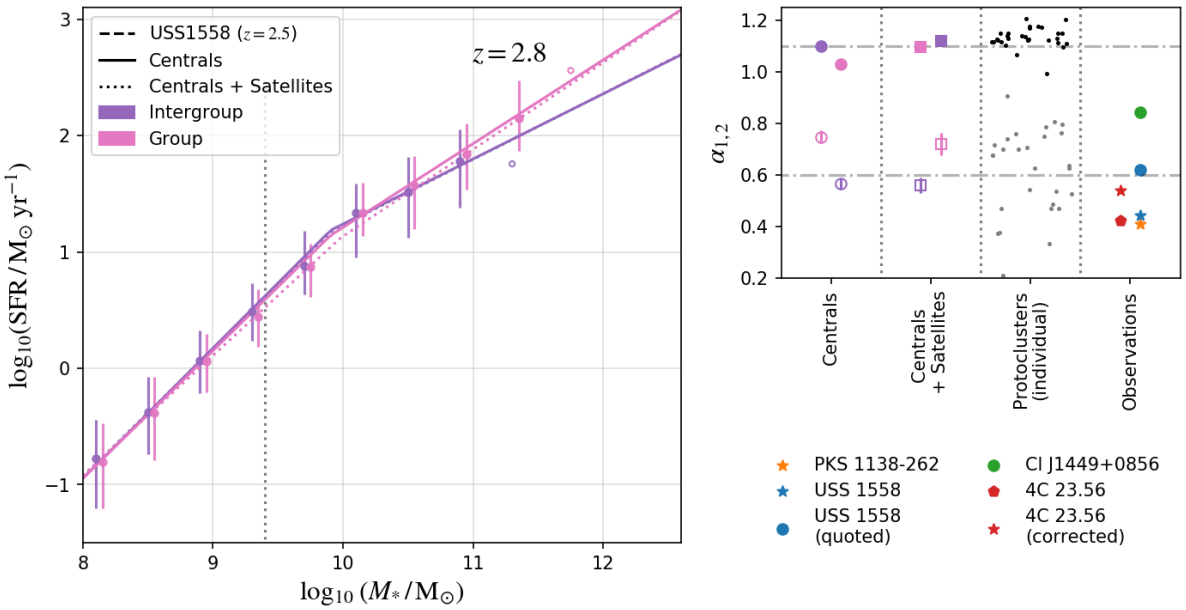

Figure 4.9: *Left panel:* the protocluster star-forming sequence at $z = 2.8$ decomposed into dense groups (pink) and intergroup (purple) populations (see criteria in Section 4.3.3). Points show the binned means with 1σ scatter; non-filled points are shown where there are few than ten galaxies in a bin. The fit relation is shown for centrals (solid lines) and centrals + satellites (dotted lines). Observational results for USS 1558 from the MAHALO survey (Shimakawa et al., 2017a) are shown by the dashed lines, for a fixed gradient $m = 0.62$. *Right panel:* high- and low-mass gradient for the group and intergroup regions. Also shown are individual protocluster fits in grey and black (high- and low-mass respectively) and observational results estimates from Shimakawa et al. (2018, 2017a); Tanaka et al. (2011); Smith et al. (2019). In the simulations, galaxies in dense groups above the turnover mass exhibit higher star formation rate than those in the intergroup population, showing a similar offset to that seen in the observations.

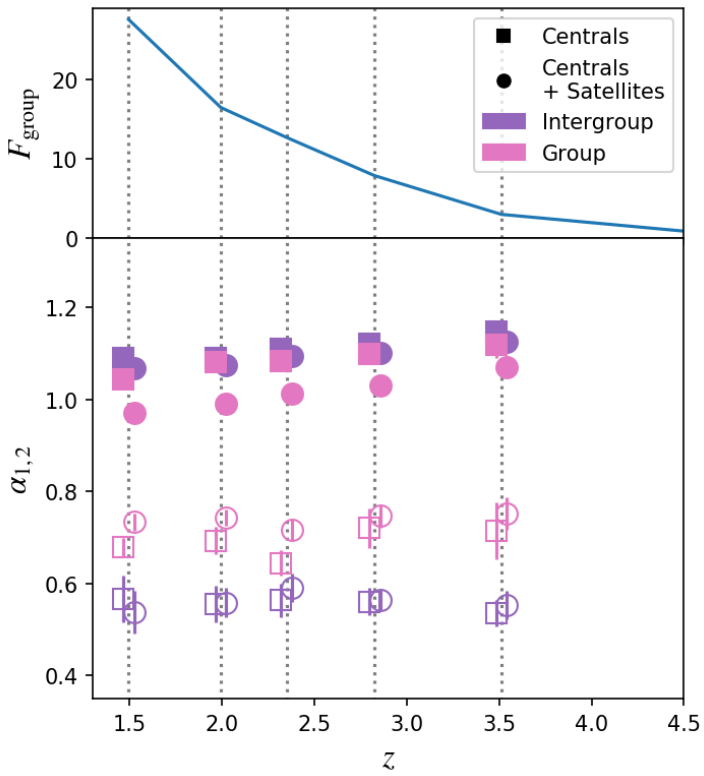

Figure 4.10: *Top:* the fraction of protocluster galaxies in groups against redshift. *Bottom:* the group (pink) and intergroup (purple) high- and low-mass slope against redshift.

distributions, which appear to show similar behaviour between the group and intergroup populations. The group galaxy relation extends to higher masses than the intergroup, which reflects the enhanced clustering around high mass halos. We fit the group and intergroup distributions as in Section 4.3.1, and evaluate the uncertainties with a non-parametric bootstrap analysis. The right panel of Figure 4.9 shows the fit for the high- and low-mass slope at $z = 2.8$. The low mass slope is similar in both environments ($\alpha_1 \sim 1.0$), but above the turnover mass group galaxies exhibit a significantly steeper SFS ($\alpha_2^{\rm group} \sim 0.75$; $\alpha_2^{\rm intergroup} \sim 0.58$), which translates to higher star formation rates. This qualitatively matches the behaviour seen in USS1558 by Shimakawa et al. (2017a), where they see an offset in the normalisation of ~ 0.15 dex between group and intergroup populations, though their measured slope is shallower.

In Figure 4.10 we show the redshift evolution of the fit parameters in the group and intergroup populations. The group high-mass slope remains significantly (outside the bootstrap uncertainties) above the intergroup at all redshifts ($\alpha_2^{\rm group} \sim 0.7$; $\alpha_2^{\rm intergroup} \sim 0.55$), both including and excluding satellites.

The low-mass slopes show more similarity between the group and intergroup populations, except for the group relation including satellites. This has a lower slope at all redshifts, and the offset gets marginally larger with redshift ($\Delta\alpha_2 = 0.06$ at $z = 3.5$; $\Delta\alpha_2 = 0.1$ at $z = 1.5$). This shows that the offset in the centrals+satellite relation for the low-mass slope (shown in Figure 4.4) is predominantly due to the group satellite population.

4.3.4 Differences between the protocluster & field SFS

Our results so far paint a picture of subtle but significant differences between the SFS in protoclusters and the field, as well as diversity between protoclusters in the high-stellar mass regime. We have also seen the influence of dense groups within protoclusters, which promote star formation in high-mass centrals, whilst inhibiting star formation in low-mass satellites.

To formally evaluate these differences we apply the Kolmogorov-Smirnov test to the stellar mass, SFR and sSFR cumulative distribution functions, for star-forming centrals. We perform the test between the protocluster and field populations, the group and intergroup populations, and finally the intergroup and field populations. We bootstrap the test and take the median statistic to prevent bias as well as sensitivity to high-mass outliers. We also perform the test independently on the high- and low-mass regimes, split using the fit turnover mass, to reveal any mass-dependent trends. Tables 4.1, 4.2 and 4.3 summarise the p-values returned for each test.

The protocluster and field stellar mass and SFR distributions are significantly discrepant ($p \ll 0.05$) at all redshifts. However, for the sSFR distribution there is no significant discrepancy ($p > 0.05$) except at $z = 6.7$. This suggests that, except at the highest redshifts, protocluster galaxies have higher stellar masses and star formation rates on average than those in the field, but do **not** have higher SFR at fixed stellar mass. Breaking this down in to high- and low- mass regimes, we see that the sSFR distribution is not discrepant in the low-mass regime, but does show a significant discrepancy in the high-mass regime at all redshifts except $z = 4.61$. These results support the significantly different fits to the high-mass slope between protoclusters and the field, shown in Figure 4.4.

How much do the groups within protoclusters drive this discrepancy? The intergroup sSFR distributions show no significant discrepancy with the field, except at $z = 6.7$.

	Protocluster - Field			Group - Intergroup			Intergroup - Field		
z	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$
1.49	0.466	0.466	0.0	0.010	0.219	0.0	0.370	0.370	0.031
1.99	0.433	0.252	0.0	0.003	0.134	0.0	0.301	0.200	0.013
2.35	0.466	0.370	0.0	0.181	0.466	0.0	0.466	0.370	0.005
2.83	0.433	0.370	0.003	0.005	0.121	0.0	0.401	0.341	0.097
3.51	0.097	0.069	0.022	0.008	0.121	0.0	0.073	0.062	0.055
4.61	0.069	0.087	0.164	0.0	0.008	0.0	0.058	0.078	0.200
6.77	0.020	0.048	0.033	0.0	0.0	0.0	0.015	0.055	0.020

Table 4.1: Kolmogorov-Smirnov test p-value results for the cumulative sSFR distributions. Results are shown between protocluster and field, group and intergroup and field and intergroup populations. The p-values are computed from the median of the KS-statistic from a bootstrap analysis on each population.

	Protocluster - Field			Group - Intergroup			Intergroup - Field		
z	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$
1.49	0.006	0.114	0.0	0.006	0.156	0.0	0.055	0.190	0.219
1.99	0.001	0.048	0.0	0.0	0.370	0.0	0.008	0.048	0.370
2.35	0.001	0.038	0.0	0.0	0.341	0.0	0.003	0.048	0.048
2.83	0.0	0.013	0.0	0.0	0.263	0.0	0.0	0.010	0.011
3.51	0.0	0.001	0.0	0.0	0.108	0.0	0.0	0.001	0.0
4.61	0.0	0.002	0.0	0.0	0.0	0.0	0.0	0.001	0.001
6.77	0.0	0.001	0.0	0.0	0.0	0.0	0.0	0.001	0.0

Table 4.2: As for Table 4.1, but showing results for the SFR distributions.

	Protocluster - Field			Group - Intergroup			Intergroup - Field		
z	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$	total	$M_* < M_{*,0}$	$M_* > M_{*,0}$
1.49	0.005	0.078	0.0	0.002	0.087	0.0	0.038	0.148	0.026
1.99	0.002	0.062	0.0	0.0	0.433	0.0	0.010	0.078	0.004
2.35	0.001	0.026	0.0	0.0	0.341	0.0	0.001	0.033	0.0
2.83	0.0	0.007	0.0	0.0	0.288	0.0	0.0	0.006	0.0
3.51	0.0	0.002	0.0	0.0	0.087	0.0	0.0	0.002	0.0
4.61	0.0	0.002	0.0	0.0	0.000	0.0	0.0	0.002	0.0
6.77	0.0	0.0	0.0	0.0	0.000	0.0	0.0	0.0	0.0

Table 4.3: As for Table 4.1, but showing results for the M_* distributions.

In contrast, the group and intergroup distributions do show significant discrepancies at almost all redshift. Breaking down in to the two mass regimes, we see that it is groups that drive the majority of the discrepancy in the high-mass regime.

In summary, dense groups within protoclusters not only extend the SFS to higher masses and SFRs, but also promote higher SFR at fixed stellar mass, particularly at the high mass end.

4.3.5 The observed protocluster star-forming sequence

So far we have limited our comparison with the observed SFS to field measurements. Unfortunately, there are few comprehensive studies of the SFS in protocluster environments. This is due to a number of factors, principally their rarity and large volume, which makes identifying and surveying them in their entirety observationally expensive. It is also difficult to distinguish true protocluster galaxies from nearby field contaminants, particularly in the absence of spectroscopic confirmation, which is often the case for low-SFR galaxies; for studying the star-forming sequence this is less of an issue, since line emission from star forming regions provides ample targets for narrowband or spectroscopic observations. Finally, those studies that do claim to study protocluster candidates often probe very different scales, from dense subgroups on ~ 100 kpc scales, up to the whole protocluster on ~ 10 cMpc scales.

Despite these difficulties, a few protocluster candidates have recently been studied in detail. Figure 4.11 shows galaxies on the SFS from four well studied protoclusters between $1.5 < z < 2.5$ (clockwise from top left): USS 1558 (Shimakawa et al., 2017a), PKS-1138 (Shimakawa et al., 2018); 4C 23.56 (Tanaka et al., 2011); and Cl J1449 (Smith et al., 2019).

4.3.5.1 PKS 1138 and USS 1558

Using narrowband imaging, the MAHALO survey selected Hα-emitters (HAEs) in USS-1558 (Shimakawa et al., 2017a). The dense substructure covers an area ~ 8 cMpc in diameter, which leads to high completeness and purity of the protocluster galaxy population in numerical studies (*i.e.* little contamination from neighbouring field galaxies, Lovell et al., 2018). The slope from a linear fit above the mass completeness limit is

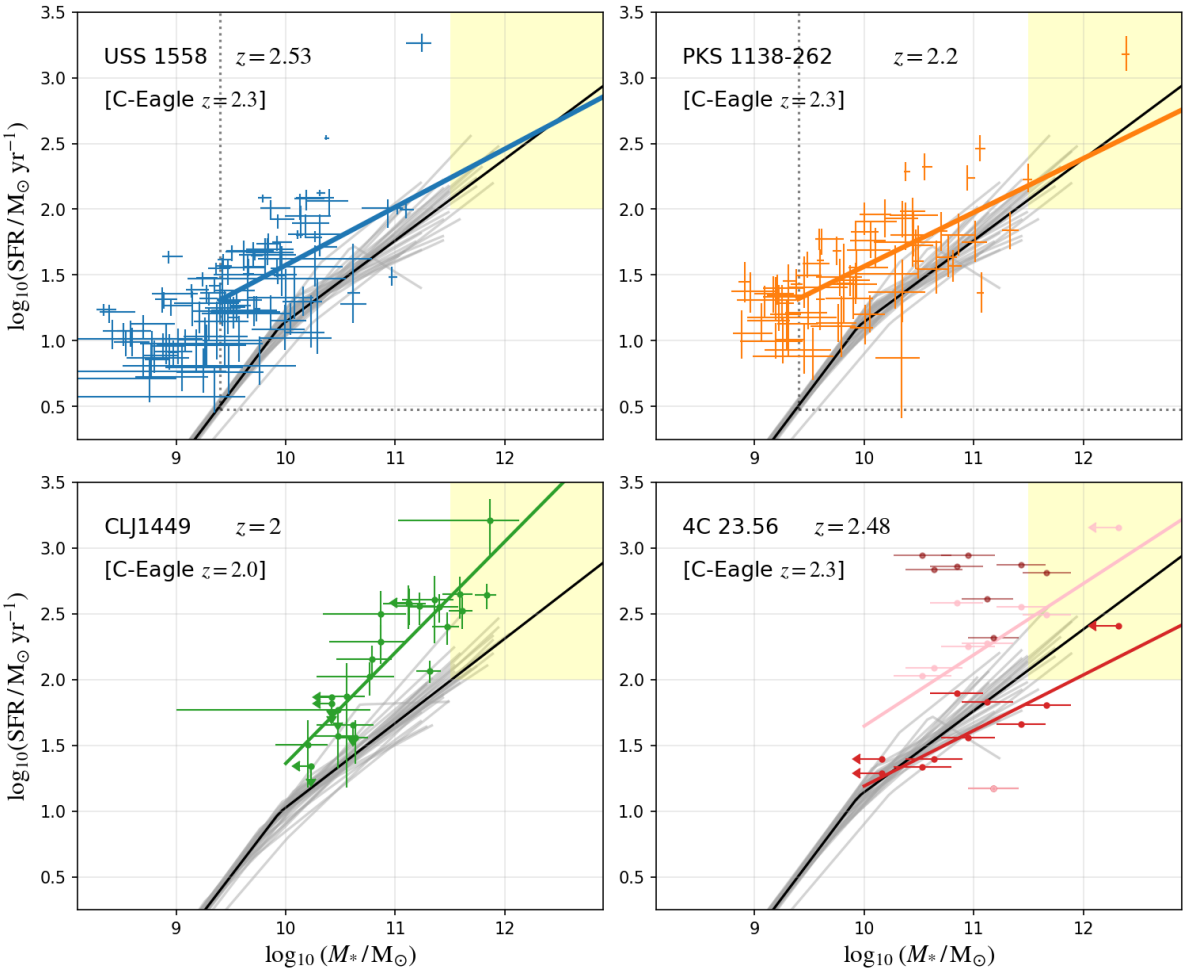

Figure 4.11: The star-forming sequence in observed protoclusters (coloured points) compared to the *C-Eagle* relation for all protocluster combined (black line) and each individually (grey lines) at the nearest redshift. *Clockwise from top left:* USS 1558 (Shimakawa et al., 2017a), PKS-1138 (Shimakawa et al., 2018); 4C 23.56 (Tanaka et al., 2011) (red, pink and dark red points show the Hα intrinsic, Hα dust corrected and Spitzer MIPS-based SFR estimates); and Cl J1449 (Smith et al., 2019). The dashed lines shows the approximate survey stellar mass and SFR completeness limits, where provided.

shallow ($\alpha = 0.44$), but within the range probed by our protocluster sample. However, the normalisation at the turnover mass is significantly higher than in C-EAGLE, greater than the discrepancy seen in the field. This may be due to the Kennicutt (1998) prescription for the Hα-SFR calibration used; Hayashi et al. (2016) suggest that SFR derived from Hα may be overestimated if metallicities are lower than typically assumed from the mass-metallicity relation. The contribution of binaries may also lead to biases in the assumed calibration; using binary population synthesis could ameliorate the discrepancy (e.g. BPASS, see Wilkins et al., 2019; Stanway & Eldridge, 2018). Interestingly, Shimakawa et al. (2017a) highlight star-bursting behaviour in intermediate mass galaxies as a possible cause; in Section 4.3.6 we show that the scatter around the SFS for galaxies at the turnover mass is significantly higher in protoclusters, so this may be a contributing factor.

The same MAHALO survey also studied PKS-1138, also known as the 'spiderweb galaxy' (Shimakawa et al., 2018), and found similar behaviour to USS 1558: a shallow slope ($\alpha = 0.41$) and positive offset in normalisation at the turnover. Shimakawa et al. (2018) note that $E(B-V)_{\rm stellar} = E(B-V)_{\rm nebular}$ is assumed, which may not be the case for the most highly star-forming galaxies at $z \sim 2$ (Price et al., 2014; Reddy et al., 2015); this would lead to *underestimated* SFRs for these galaxies, giving a steeper measured slope from both studies, and increasing the discrepancy with our simulations.

4.3.5.2 Cl J1449+0856

Smith et al. (2019) measured the dust-obscured star formation in Cl J1449+0856, a well studied protocluster at $z = 2$ (Coogan et al., 2018; Strazzullo et al., 2018). They use SCUBA-2 and Herschel observations combined with optical ancillary data to perform SED fitting with CIGALE, and derive star formation rates and stellar masses. The normalisation of the observations is higher as seen before, but the slope of the relation is steeper compared to the MAHALO results, and in better agreement with the high-mass slope. In fact, the slope is very similar to that derived in the dense groups within the protocluster (see Figure 4.9, discussed in Section 4.3.3).

4.3.5.3 4C 23.56

Tanaka et al. (2011) present SFR and stellar mass estimates of Hα selected galaxies in 4C

using BC03 models assuming a Chabrier IMF

23.56, though only 4 candidates are spectroscopically confirmed. The aperture on the sky is ~ 4.1 cMpc, however the redshift uncertainty for the unconfirmed candidates means that the probability of field contamination along the line of sight could be high. Both the original and dust-corrected SFR estimates have similar, shallow slopes ($\alpha = 0.42$ and 0.54, respectively), though the latter has a higher normalisation, as already discussed. Tanaka et al. (2011) suggest that PKS-1138 is in a more evolved state than 4C 23.56, since the latter is still vigorously forming stars. This may be due to the high gas density measured in 4C 23.56 (Lee et al., 2017).

4.3.5.4 Discussion

Together, these observations paint a picture of significant diversity in the slope and normalisation of the protocluster SFS between $1.5 < z < 2.5$ that qualitatively matches the diversity seen in our simulated protocluster population. The lack of dependence on descendant mass suggests a different source for this diversity than the total mass and volume, and the observations suggest this may be due to smaller scale group environments within the overall web-like structure of the protocluster. In Section 4.3.3 we investigate this by decomposing our simulated protoclusters into group and intergroup populations.

4.3.5.5 Proto-BCG galaxies

The high mass, high SFR galaxies at the top right of the star-forming sequence, such as the spiderweb galaxy itself in PKS 1138-262, are often assumed to be proto-BCGs. In C-Eagle we find a high fraction of proto-BCG objects in this region of the parameter space, 80% of objects at $z \sim 2.3$ (where $M_* \,/\, M_\odot > 10^{11.5}$ and $\log_{10}(SFR \,/\, M_\odot \text{ yr}^{-1}) > 2$, shown by the yellow square in each panel of Figure 4.11). Whilst this fraction is high, it highlights a non-negligible population of massive, star-forming galaxies in protoclusters that do *not* have a BCG descendant.

4.3.6 Scatter in the Star-Forming Sequence

The scatter in the star-forming sequence contains information on the variability in the recent star formation history of galaxies at a given mass (Matthee et al., 2017; Katsianis et al., 2019). A tight relation implies smooth, accretion driven growth at high redshift, whereas greater scatter would suggest less coherent, bursty stellar mass growth across the

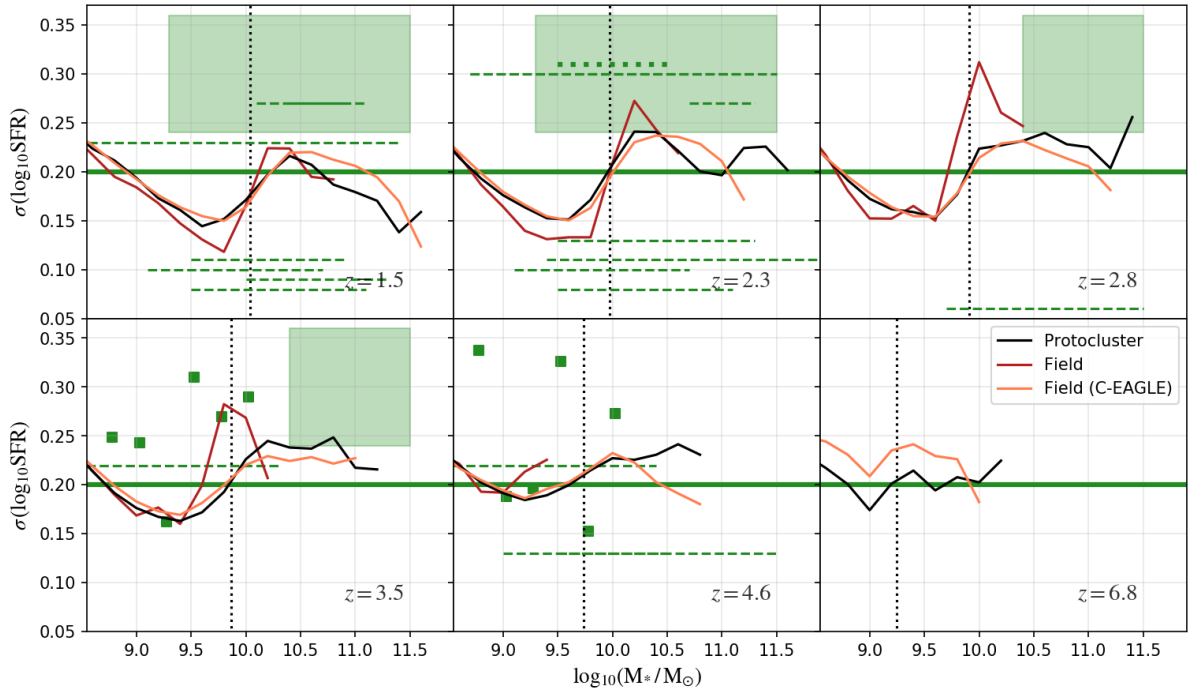

Figure 4.12: 1σ scatter around the star-forming sequence for central galaxies in protoclusters (black), the field (red) and the C-EAGLE field region (orange) between redshifts $z = 1.5 - 7$. The scatter is measured around the best-fit piecewise relation measured in Section 4.3.1 for each population. The combined, mass- and redshift-independent intrinsic scatter from Speagle et al. (2014) is shown (bold green, 0.2 dex), as well as individual measurements from this study at their respective redshift and stellar mass ranges in each panel (dashed green). We also show results from Schreiber et al. (2015) (green filled region) Shivaei et al. (2015) (green dotted) and Salmon et al. (2015) (green squares).

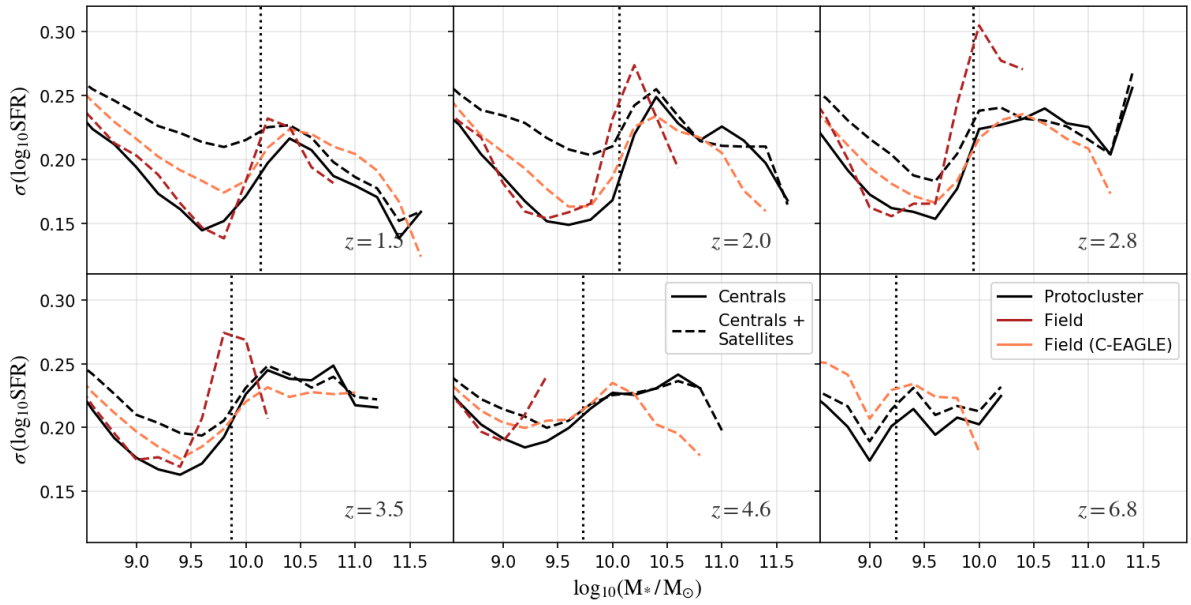

Figure 4.13: As for Figure 4.12, but including centrals *and* satellites (dashed). The centrals only protocluster relation is shown for comparison (black, solid). Note that the y-axis limits have been changed from Figure 4.12 for clarity.

galaxy population, possibly through mergers. Underestimating the scatter would suggest that the conversion of accreted gas into stars in the model is too smooth, whereas an overestimate in the scatter could be the result of feedback processes being too strong or stochastic. The scatter is obviously sensitive to the timescale of the SFR indicator; here we use the instantaneous SFR taken from the dense star-forming gas, but using a longer timescale SFR indicator would smooth out shorter episodes of variation.

4.3.6.1 The scatter in the centrals-only relation

We measure the 1σ scatter around the best fit two-part piecewise relation measured in Section 4.3.1 in bins of stellar mass, after implementing the sSFR cut for passive galaxies. Figure 4.12 shows the evolution of the stellar-mass dependent scatter in the star-forming sequence for centrals only. The general trend, for all models, environments and redshifts, is that the scatter tends to decrease from $\log_{10}(\mathrm{M}_* / \mathrm{M}_\odot) = 8.5$ to some characteristic mass ~ 9.5. Above this it increases dramatically and plateaus. This mass at which the scatter increases is similar to the turnover mass in each model and environment, shown (for the protocluster fit) as the vertical dotted lines in each panel. As the turnover mass evolves to lower masses with increasing redshift, so does the mass at which the scatter increases.

The higher scatter for very low mass galaxies ($\log_{10}(\mathrm{M}_* / \mathrm{M}_\odot) < 9$) can be explained by a scenario where feedback from star formation, which dominates in this mass regime, is more effective at expelling gas from lower mass galaxies due to the shallower potential. Higher mass galaxies tend to be able to retain their gas reservoirs despite energetic supernovae feedback. The increase in the scatter around the turnover mass has been attributed to the onset of AGN feedback (Matthee et al., 2017), visible in the field region at $z < 3.5$ and in the protocluster regions at even higher redshifts due to the greater number of high mass galaxies.

Recently, Katsianis et al. (2019) studied the scatter in the SFS as a function of stellar mass in the periodic EAGLE simulations up to $z = 4$. They use a similar sSFR cut to Matthee et al. (2017), and found similar behaviour to what we find in the protocluster environment: an increase in the scatter at high stellar masses attributable to AGN feedback, and an increase at lower masses attributed to efficient stellar feedback. We are able to extend the relation to higher stellar masses due to the large galaxy sample from the protoclusters, and find that the scatter remains high at $M_* / M_\odot > 10^{11.5}$ for $z > 2$, but falls above this mass at $z \leqslant 2$. This may be due to reduced AGN feedback during this phase of collapse; Figure 4.5 shows that the ratio of black hole mass to halo mass is lower for the most massive galaxies at $z = 2$ compared to higher redshifts.

Interestingly, Katsianis et al. (2019) find no dependence of the scatter on recent mergers. We expect this to be more common in the dense protocluster environment, but also see no environmental difference for centrals. However, mergers are not the only environmental effect in dense environments: the large number of high mass galaxies and overall volume density of galaxies means that there are a large number of satellites of a range of masses. In the next section we study the scatter including satellites in protoclusters and the field.

4.3.6.2 The satellite-induced scatter

Figure 4.13 shows the scatter in the star-forming sequence including both centrals *and* satellites. In all regions, the scatter is similar above the turnover mass when including satellites, but below the turnover mass there is significant environmental dependence. Low mass satellites in protoclusters lead to a flattening of the scatter as function of stellar mass by $z = 1.5$ (+0.07 dex at $M_* / M_\odot = 10^{9.5}$), wherease in the C-EAGLE

field region the increase is less dramatic (+0.03 dex). This is expected in the dense protocluster environment due to the effects of interactions and mergers on the short-timescale star formation rate. Perhaps more surprising is that the increased low-mass scatter in protoclusters is present up to $z \sim 4.5$, which suggests the environmental effect on the star formation histories of satellites is present at very early times.

4.3.6.3 The observed intrinsic scatter

The observational scatter is the product of the intrinsic scatter convolved with the evolution of the star-forming sequence in the measurement redshift bin (Noeske et al., 2007), as well as being sensitive to uncertainties in redshift, stellar mass and star formation rate measurements (Speagle et al., 2014; Katsianis et al., 2019). Eddington bias can also affect the measure of scatter, since low-mass galaxies with, on average, lower SFR will be up-scattered into higher mass bins (Speagle et al., 2014). The SFS scatter is also sensitive to the threshold for quiescence used; including more quiescent objects will increase the scatter significantly, as expected (Katsianis et al., 2019). However, we note that at low redshift, where these effects are smaller and the experimental contribution better constrained, the intrinsic scatter in the model matches observational constraints (Matthee & Schaye, 2019).

A number of observational studies have attempted to derive the intrinsic scatter at high redshift; we show a selection in Figure 4.12. Speagle et al. (2014) combine a number of observational studies, measuring a stellar-mass independent intrinsic scatter of 0.2 dex at all redshifts, shown as the bold green line in each panel. Figure 4.12 also shows a number of the individual relations from Speagle et al. (2014) as the dashed green lines, at their respective redshift and mass ranges. We also show a number of more recent measurements of the intrinsic scatter (Schreiber et al., 2015; Shivaei et al., 2015; Salmon et al., 2015). It is clear from the observations that at high redshift there is considerable inter-study scatter, and that our simulation results lie within this scatter.

It is interesting to note the apparent bimodality of the observations, with some studies predicting intrinsic scatter between 0.2-0.35 dex, and others predicting much lower intrinsic scatter, around 0.05-0.15 dex. This may be due to some observations probing the post-turnover scatter, whilst others find more galaxies at the turnover, where the scatter is at

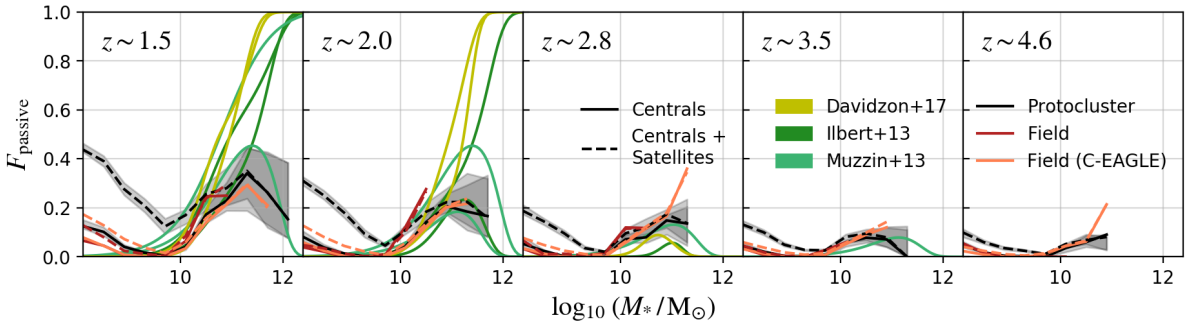

Figure 4.14: Evolution of the passive fraction from $1.5 < z < 4.6$ for protoclusters (black), field (red) and C-EAGLE field (orange) regions. Passive galaxies are defined as those whose SFR is lower than the sSFR cut at a given redshift. The relation is shown where there are $\geqslant 10$ objects per bin. Solid lines show the relations for centrals only, dashed lines when including satellites. The effect of using a higher- and lower-sSFR cut on the protocluster passive fraction is shown by the shaded grey region. Observed field relations from Davidzon et al. (2017), Muzzin et al. (2013) & Ilbert et al. (2013) are shown in green; where the redshift range of the simulation lies between the observations both the upper and lower redshift observational constraints are plotted.

a minimum. However, the observations do not show a strong dependence on stellar mass range for whether they lie in the high- or low-scatter regime. When including satellites, the stellar-mass dependence disappears at $z \sim 2$ in protocluster environments. This may explain the stellar-mass invariance in some observational studies that are large enough to include such overdense environments. Further observational results, that are capable of constraining the true intrinsic scatter, are required to determine if the scatter about the star-forming sequence is indeed stellar-mass dependent, and the effect of satellites in overdense environments.

In summary, we find that the scatter about the SFS is highly stellar mass dependent in both field and protocluster environments, showing an increase at the turnover mass associated with the onset of AGN feedback, and a large increase at lower masses in protocluster environments when including satellites (for $z \leqslant 4.5$).

4.4 Passive fractions

In the previous sections we have characterised the behaviour of the star-forming galaxy population in protoclusters. We now investigate the passive galaxy population, and whether it shows any environmental dependence.

Figure 4.14 shows the evolution in the passive fraction, defined as all galaxies whose SFR

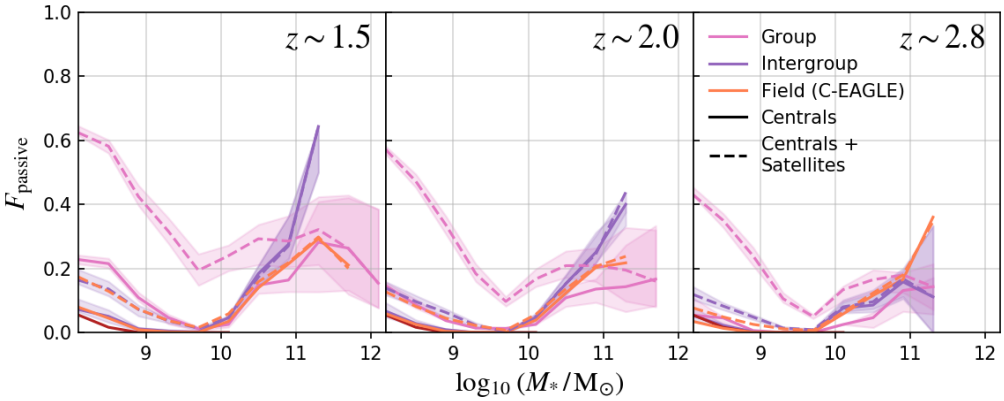

Figure 4.15: Passive fraction in protoclusters split into group (pink) and intergroup (purple) populations, along with the relation in the periodic field regions (red) and the C-EAGLE field regions (orange), split in to centrals only (solid) and centrals + satellites (dashed). The relations are plotted where there are greater than 10 galaxies in a given bin.

lies below the sSFR cut at that given redshift (shown as the grey points in Figure 4.3). The general trend across all environments is for the passive fraction to decrease toward the turnover mass, as stellar feedback becomes less effective, then increase above this where AGN feedback dominates. Considering centrals only, there is no significant environmental dependence up to $z = 1.5$. At $z = 1.5$ the protocluster passive fraction is higher than the field by ~ 0.1 dex in both the low- and high-mass regimes. This suggests environmental quenching of high and low-mass centrals becomes effective in the latter collapse stage of the cluster, but not at higher redshifts.

Figure 4.14 also shows observational constraints from Davidzon et al. (2017), Muzzin et al. (2013) & Ilbert et al. (2013), derived by comparing Schechter fits to the stellar mass function for active and quiescent populations. The agreement above the turnover mass of the simulations and the observations is reasonably good, though Davidzon et al. (2017) predict higher passive fractions than the other two studies.

Including satellites leads to a more dramatic dependence on protocluster environment. The low-mass normalisation is higher than the field at all redshifts, from 0.05 dex at $z = 4.6$ to 0.25 dex by $z = 1.5$. To test whether this is dependent on dense groups within the protocluster, we once again study protocluster galaxies split into group and intergroup populations according to the algorithm detailed in Section 4.3.3. Figure 4.15 shows the passive fractions in groups and intergroup from $z = 1.5 - 2.8$. Below the turnover mass it is clear that it is the dense group environments that dominate the passive satellite

fraction. The intergroup passive fraction including satellites is only marginally higher than the centrals only (< 0.05 dex at all redshifts). Satellites in dense groups are more efficiently quenched, most likely through interactions with the central and other satellites in the congested group environment.

Interestingly, at $z \leqslant 2$, the passive fraction at the high-mass end in the intergroup region is *higher* than in both the group and field regions, for both centrals only and when including satellites. This suggests that intergroup galaxies, despite showing no significant discrepancy with field galaxies in their sSFR distributions along the SFS at these redshifts (see Section 4.3.4), have higher quenched fractions than their field and group counterparts at the same mass. A possible explanation could be that the intergroup environment is still denser than the field when normalised by volume, and so interactions are more likely, which could lead to quiescence. Another explanation is that massive intergroup galaxies experience a kind of 'strangulation', whereby accretion of cold gas is disrupted by proximity to dense groups. Whilst such interactions are equally, if not more likely, in the groups, the positive effect of abundant cold gas in the groups could end up overriding the detrimental effect of interaction, promoting star formation rather than inhibiting it.

It is difficult to distinguish passive populations in protoclusters, since their redshifts are primarily determined photometrically, which incurs large uncertainties, making it difficult to confirm their protocluster membership. Despite this, their are some constraints on the passive fraction in a number of $z \sim 2$ protocluster candidates (Lee-Brown et al., 2017; Cooke et al., 2016; Newman et al., 2014). Figure 4.16 shows the passive fraction in these protoclusters, alongside field constraints published in these studies, and also the dedicated field constraints from Davidzon et al. (2017), Muzzin et al. (2013) & Ilbert et al. (2013). The observed protocluster passive fraction is higher than in the field at all measured masses, and rises to unity by $M_* \,/\, M_\odot \sim 10^{10.5}$ in all three observed protoclusters.

We caution that such measurements are notoriously difficult; Shimakawa et al. (2018) found a passive fraction in PKS-1138 of $\sim 36\%$, though they note that the uncertainties are very large, making any inferences difficult. With this in mind, we tentatively conclude that, even for the most lenient sSFR cut, we do not predict protocluster passive fractions within a factor of 2 of those observed. We also see a decrease in the passive fraction for the most massive ($M_* \,/\, M_\odot > 10^{11.5}$) galaxies that is not seen in the observed protoclusters,

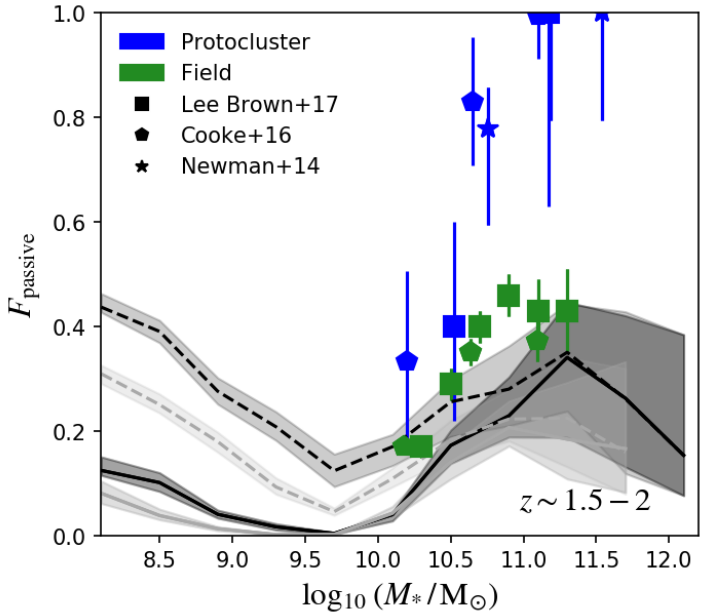

Figure 4.16: Protocluster passive fraction at $z = 1.5$ (black) and $z = 2$ (grey) for centrals (solid) and centrals + satellites (dashed). Observed protocluster passive fractions from Lee-Brown et al. (2017), Cooke et al. (2016) & Newman et al. (2014) are shown (blue points), along with any comparison field measurements where available (green points). The field relations from Davidzon et al. (2017), Ilbert et al. (2013) & Muzzin et al. (2013) are also plotted (green lines). The passive fraction in observed protoclusters is higher than the field at $M_* \,/\, M_\odot > 10^{10}$, reaching unity at $M_* \,/\, M_\odot \sim 10^{11}$, an environmental dependence we don't see in the simulations.

but is tentatively suggested by field measurements It is interesting to note that these high passive fractions are similar to those in the protocluster intergroup regions, which suggests the observed passive galaxies may reside in between the dense groups.

4.5 Discussion

4.5.1 The offset in normalisation of the star-forming sequence at cosmic noon

The star-forming sequence at $z \sim 2$ measured in the simulations has a normalisation $\sim +0.3$ dex lower than that measured in observations. This was noted in Furlong et al. (2015) as an offset in the sSFR-stellar mass relation, and attributed to a lack of bursty star formation, which would lead to higher specific star formation rates without affecting the global average stellar density evolution. However, this discrepancy is not unique to EAGLE, and has been noted by a number of authors (Davé, 2008; Sparre et al., 2015; Davé et al., 2019). What is remarkable is the consistency with which different simulations, both semi-analytic and hydrodynamic, employing very different subgrid physics recipes, predict a similar offset (Katsianis et al., 2016). The observations, in contrast, show contradictory behaviour, particularly between different observational tracers (Katsianis et al., 2017).

There are a number of possible remedies to this discrepancy. A factor could be that stellar populations in high-z galaxies emit harder ionising radiation, however this is also expected to be the case at redshifts above $z \sim 2$ where we do not see a strong discrepancy. As discussed in Section 4.3.5 the Hα-SFR calibration may be affected by the inclusion of binary stellar populations; using a more conservative calibration would reduce the estimated SFR, reducing the tension between Hα measures of the star-forming sequence at $z \sim 2$ (Wilkins et al., 2019; Stanway & Eldridge, 2018). Observational results for the stellar mass density and cosmic star formation rate density are also in tension during this key epoch of peak star formation rate density (Wilkins et al., 2008; Yu & Wang, 2016), an offset that has been attributed either to an evolving IMF at high redshift, or over/under-predictions of the observational SFR or stellar mass, respectively, both of which would also address the discrepancy in the star-forming sequence fit. Using modern SED fitting approaches, Leja et al. (2018) found a simultaneous increase in the estimated

stellar mass as well as a drop in estimated SFR, leading to a 0.3 dex reduction in the sSFR, which would address the discrepancy seen here and in other models. It remains to be seen whether a similar approach at $z > 2.5$ will preserve the good agreement with the simulations.

4.5.2 The effect of the protocluster environment on the star-forming sequence

The C-EAGLE simulations show great diversity in the form of the star-forming sequence in protoclusters, that is dependent not on the descendant cluster mass, but on the presence of dense groups within the overall protocluster superstructure. This predominantly affects the slope and normalisation at the high-mass end. The high normalisation of the star-forming sequence in USS-1558 has been tentatively attributed to starbursts in intermediate mass galaxies (Shimakawa et al., 2017a). Cibinel et al. (2019) find a higher merger fraction for starburst galaxies above the star-forming sequence, which begs the interesting question of whether these starbursting objects are more common in the protocluster environment. We leave a detailed investigation of merger fractions and their impact on the star-forming sequence to future work.

The protocluster environment also has a significant impact on the satellite galaxy population. This is evidenced by the 0.3 dex higher scatter in the SFS for satellites below the turnover mass, and the higher passive fractions of satellites below and around the turnover mass. There are a number of physical processes in dense environments that could be responsible for this behaviour. The probability of mergers and interactions is higher, however this is also the case for centrals, where there is little difference with the field population up to $z \sim 2$. The large gas mass in the IGM in protoclusters available for accretion could lead to both higher SFRs and higher accretion on to the central SMBH. The latter could lead to more energetic feedback events and quenching.

4.5.3 Brightest Cluster Galaxy masses

Bahé et al. (2017) found that the $z = 0$ stellar mass of BCGs in C-EAGLE is higher than the observed relations, a discrepancy that is positively correlated with cluster mass. However, the total stellar mass in the clusters is in agreement with observational constraints. The

question is, why is BCG assembly so efficient in C-EAGLE; are too many stars formed *in-situ*, or are too many stars assembled through mergers? We do not attempt to answer this question explicitly in this paper, however we can compare the proto-BCG evolution along the SFS at high-z to provide some clues. The majority of these high mass, high SFR galaxies are BCG progenitors (see Section 4.3.5.5), so we can think of the most massive as being the *main-branch* progenitor, and host to *in-situ* star formation, whereas star formation in other progenitors is then *ex-situ*.

The most massive, highly star forming objects in the simulated protoclusters are of comparable mass and SFR to those in PKS 1138, USS 1558, 4C 23.56 and Cl J1449 (Shimakawa et al., 2017a, 2018; Tanaka et al., 2011; Smith et al., 2019). This on its own suggests that the assembly of stellar mass into the BCG progenitor main branch is not too efficient up to $z \sim 2$, at least compared to this heterogeneous observational sample. Bahé et al. (2017) found that only $\sim 10\%$ of the BCG stellar mass was formed at $z < 1$, which suggests that there is either significant star formation in BCG main progenitors between $1.0 < z < 1.5$, or too many stars are formed *outside* the main branch of the BCG progenitor merger tree, that are then accreted later. We will investigate the formation and assembly of stellar mass in clusters and their BCGs explicitly in future work.

4.5.4 Selection biases and future surveys

Galaxy protoclusters have been identified through a range of different techniques and tracers, leading to a heterogeneous sample of candidates. This can lead to a number of biases, both in the selection of the protocluster sample as well as the characterisation of the galaxy population used to measure the SFS.

AGN have been proposed as potential tracers of overdense environments, since these overdense environments will contain both supermassive black holes as well as abundant gas for accretion. The Clusters Around Radio-Loud AGN (CARLA) survey discovered a number of clusters and protoclusters between $1.3 < z < 3.2$ using HzRGs (Wylezalek et al., 2013; Cooke et al., 2015). Both PKS 1138 and USS 1558 were first identified from targeted follow up of their central radio galaxies (Pentericci et al., 2000; Kajisawa et al., 2006), However, protoclusters identified through such objects may represent a biased sample compared to the total protocluster population. Cosmological models have also revealed a

complex relationship between AGN and galaxy overdensities (Orsi et al., 2016; Habouzit et al., 2018). We will investigate the coincidence of AGN and protocluster environments in the C-Eagle sample in future work.

A number of ongoing surveys have discovered large numbers of protocluster candidates, that open up the possibility of follow up with future observatories to characterise the SFS. The Subaru/Hyper Suprime-Cam (HSC) has been a workhorse for protocluster studies over recent years: the SILVERRUSH program has discovered large numbers of protocluster candidates up to redshifts of $z \sim 7$ though wide field surveys of Lyman-α emitters (Higuchi et al., 2018; Harikane et al., 2019); and the GOLDRUSH program has discovered a number of $z \sim 4$ protoclusters through the dropout technique. The MAMMOTH (MApping the Most Massive Overdensity Through Hydrogen) survey (Cai et al. 2016, Cai et al. 2017) is a novel method of identifying protocluster candidates through absorption mapping along quasar sightlines. Compilation of previous heterogenous surveys is also another promising avenue for protocluster identification and chracterisation; the Candidate Cluster and Protocluster Catalogue (CCPC) was one of the first such compilations, using a consistent surface-density criterion Franck & McGaugh (2016a,b). However, as with AGN, the exact relationship of the chosen tracer to the underlying overdensity can affect the magnitude of the measured overdensity, as well as its shape and position relative to the underlying matter overdensity (see Shi et al., 2019).

There are a number of future observatories and surveys planned that will provide improved protocluster samples, as well as follow up of known protocluster candidates to characterise the SFS and passive population in detail. Future NIR surveys from Keck/MOSFIRE and the NIRSpec instrument on JWST will be able to distinguish passive galaxies with high completeness and accurate redshifts, increasing the passive sample in protocluster environments.

4.6 Conclusions

We study the star-forming sequence in high-redshift protoclusters and the field in the Eagle simulation. Our results are as follows:

- The star-forming sequence in protoclusters and the field show similar *overall* behaviour, rising in normalisation with redshift, and exhibiting a turnover at a stellar

mass coincident with the onset of AGN feedback. The slope and normalisation are in good agreement with observational constraints, except at cosmic noon ($z \sim 2$) as seen in other numerical studies. The spread in measured observational slopes is shown to be due to the low-mass incompleteness, in many cases, of high redshift surveys, which can only constrain the shallower high-mass slope.

- Exploring the SFS fits in detail reveals significant differences between the protocluster and field environments. In the high mass-regime protoclusters have a diversity of slopes and normalisations driven by the presence of dense groups, which promote star-formation. In the low-mass regime, satellites in groups experience greater environmental harassment, reducing their star-formation.

- The turnover mass evolves to higher stellar mass with decreasing redshift, in contradiction with recent observational results. We argue that its evolution is the result of the increased efficacy of AGN feedback in lower stellar mass galaxies at high-z, which may be due to their compact morphologies and higher relative halo masses.

- We compare the star-forming sequence to a number of well studied protoclusters at $z \sim 1 - 2.5$. We find a range of high mass slopes, which may be due to both different observational tracers as well as a diversity in evolutionary stages. In the simulated protoclusters we see a similar diversity in high mass slope that has no correlation with descendant mass.

- The scatter in the star-forming sequence is $\sigma \sim 0.2$ dex at all redshifts, but shows significant dependence on stellar mass due to the competing influence of stellar- and AGN-feedback. There is no environmental dependence of the scatter for centrals, but including satellites leads to greater scatter in low mass protocluster galaxies up to $z \sim 5$, and above the turnover mass by $z \sim 2$.

- The Eagle model matches the passive fraction of galaxies in field environments reasonably well, but underestimates this fraction for high mass galaxies in protocluster environments by a factor of 2.

5 Learning the Relationship between Galaxies Spectra and their Star Formation Histories using Convolutional Neural Networks and Cosmological Simulations

Christopher C. Lovell,[1] Viviana Acquaviva,[2] Peter A. Thomas,[1] Kartheik G. Iyer,[3] Eric Gawiser,[3,4] Stephen M. Wilkins[1 17]

5.1 Introduction

We present a new method for inferring galaxy star formation histories (SFH) using machine learning methods coupled with two cosmological hydrodynamic simulations. We train Convolutional Neural Networks to learn the relationship between synthetic galaxy spectra and high resolution SFHs from the EAGLE and Illustris models. To evaluate our SFH reconstruction we use Symmetric Mean Absolute Percentage Error (SMAPE), which acts as a true percentage error in the low-error regime. We also make estimates for the observational and modelling errors. To further evaluate the generalisation properties we apply models trained on one simulation to spectra from the other. Finally, we apply each trained model to SDSS DR7 spectra, and find smoother histories than in the VESPA catalogue. This new approach complements the results of existing SED fitting techniques, providing star formation histories directly motivated by the results of the latest cosmological simulations.

A galaxy's integrated Spectral Energy Distribution (SED) contains information about countless physical properties, such as the stellar population age, mass, dust content, redshift, metallicity and star formation history (SFH). Different physical processes leave their imprint in different parts of the spectrum; the wider and more finely sampled the

[1]Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton, BN1 9QH, UK
[2]Department of Physics, New York City College of Technology, Brooklyn, NY 11201, USA
[3]Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway,
NJ 08854-8019 USA
[4]Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, NY 10010, USA

wavelength coverage, the more robust the interpretation of the various features of the SED is in terms of galaxy properties. One fundamental tool to determine the physical properties of a galaxy starting from photometric and/or spectroscopic observations is SED fitting, the procedure of iteratively comparing models to the observed galaxy SEDs (e.g. Walcher et al. 2011; Conroy 2013). Since the physical properties of the models are known, those of the data can be derived by maximizing the resemblance between data and models. The success and reliability of this method depends on the quality of the available template spectra, and the robustness of the fitting algorithm.

The field of SED fitting has seen enormous progress in the last decade (Conroy, 2013). Methods such as Markov Chain Monte Carlo have been used to efficiently explore the degeneracies associated with the large parameter space (e.g. Sajina et al. 2006; Acquaviva et al. 2011; Pirzkal et al. 2012; Acquaviva et al. 2012; Leja et al. 2017). However, one issue that has consistently emerged from these efforts is the difficulty of characterizing and constraining the star formation histories of galaxies. The spectral signatures of multiple non-coeval generations of stars can be mimicked by other physical effects, such as varying stellar metallicity, and older stellar populations with high mass-to-light ratios are easily hidden in observed spectra, an effect sometimes referred to as “outshining" (Maraston et al., 2010). It would be helpful, in Bayesian parameter estimation, to use priors to guide our exploration of very large and degenerate parameter spaces, but these are not readily available.

A wrongly reconstructed star formation history introduces significant biases in many parameters that are usually estimated through Spectral Energy Distribution fitting, such as stellar masses, stellar age indicators, dust content, and redshift (e.g. Mobasher et al. 2015; Pacifici et al. 2014; Iyer & Gawiser 2017; Leja et al. 2017). Acquaviva et al. (2015) evaluated the impact of different sources of non-algorithmic systematics on the recovered SED fitting parameters and concluded that a wrong star formation history is the most detrimental. Similarly, Iyer & Gawiser (2017) found that fitting the SFH using single stellar populations and simple functional forms (e.g. exponentially declining or constant models) leads to a bias of up to 70% in the recovered total stellar mass. Carnall et al. (2019) further demonstrated that simple parametric star formation histories impose strong priors on implied physical parameters. These introduce strong correlated biases that are

propagated through pipelines of results and used to infer key distribution functions and relations, such as the stellar mass function and the cosmic star formation rate density (Ciesla et al., 2017; Leja et al., 2019), critical for answering crucial questions in the study of galaxy formation and evolution.

One possible approach to solving this problem has been to introduce new parametrisations for the SFH that are less subject to the outshining bias (Behroozi et al., 2013b; Simha et al., 2014), or to develop parameter-free descriptions of the SFH (Tojeiro et al., 2007; Iyer & Gawiser, 2017; Iyer et al., 2019; Leja et al., 2019). Here we propose an alternative approach, using supervised machine learning algorithms to 'learn' the relationship between the SFH and the SEDs of galaxies. In contrast with SED fitting, where the SFH is built from some ensemble of simple stellar populations to maximise the resemblance in SED space, machine learning directly learns the relationship between the spectra and the entire SFH. We expect that this method will carry systematic uncertainties that are independent of those from SED fitting, so that our results will complement and strengthen the results of these approaches. Another strength of a machine learning-based approach is that the algorithm learns from the population ensemble, learning not only the correspondence between individual spectra and star formation histories, but also which star formation histories are common and which are unlikely, something that would be analogous in Bayesian parameter estimation to learning the SFH prior.

A number of recent studies have explored the effect of priors on derived SFHs in SED fitting approaches. Carnall et al. (2019) showed that parametric approaches implicitly impose a strong prior on the SFH that can lead to unrealistically tight posterior constraints on the SFR, and Leja et al. (2019) showed that even non-parametric fits are sensitive to the prior SFH distribution, particularly where the data are poor. Pacifici et al. (2013) proposed using SFHs from a semi-analytic model to generate a library of SEDs to be used in an SED fitting algorithm, and found that these simulation-motivated templates prefer symmetric or rising SFHs at intermediate redshifts ($0.2 < z < 1.4$), compared to the exponentially declining forms predicted using simple stellar populations. Finally, Wilkins et al. (2013a) show that using simulation-motivated enrichment and star-formation histories leads to more accurate stellar mass estimates from colour information only. These studies highlight the importance of the explicitly or implicitly assumed prior distribution of SFHs.

Machine learning methods are becoming an increasingly popular tool for Astronomers (Ball & Brunner, 2010; Baron, 2019). This is particularly the case where there is abundant low quality data for which expensive, higher quality data can be obtained and used for supervised training. However, a supervised machine learning algorithm is only as good as its learning sample. The main challenges to applying these techniques to measure properties such as star formation histories have been the following: assembling a sample of galaxies for which the "true" star formation history is known; and making sure that properties of the ensemble (the distribution of properties and their relationship to one another) are a fair snapshot of the real Universe. However, there has been significant recent progress from multiple independent teams on high-resolution cosmological models of galaxy evolution, which has for the first time provided the potential to test this technique (e.g. Simet et al., 2019).

Hydrodynamic cosmological simulations in particular are able to resolve stellar populations, producing realistic, high resolution SFHs by taking into account a number of effects, such as environmental interactions, mergers, and stellar and AGN feedback (Somerville & Davé, 2015). EAGLE (Schaye et al., 2014) and Illustris (Genel et al., 2014) are two recent hydrodynamic simulations that reproduce a number of key galaxy distribution functions. Both are necessarily tuned to a small number of observational constraints due to their limited resolution, which requires subgrid models to model physical processes below the simulation scale. Despite this, a number of observables not included in the tuning are simultaneously reproduced. Of interest for this study are the distributions of colours and photometric magnitudes, which are well reproduced in both models (Trayford et al., 2015; Torrey et al., 2015). The recent convergence of such detailed models with the observations, and within sufficiently large simulated volumes, has finally enabled them to be used as training sets for machine learning models.

In Section 5.2 we describe the method in detail, including an overview of the machine learning techniques (5.2.1), the simulations used (5.2.3) and our method for generating synthetic spectra (5.2.4) with SPECTACLE, a stand-alone python module for generating spectra from cosmological simulations (5.2.4).[18]. Our results when trained and tested on the simulations are presented in Section 5.3. Section 5.4 details our modelling of the uncertainty contribution from the observational and modelling sources. We then

[18]https://github.com/christopherlovell/spectacle

apply our trained models to SDSS observations: Section 5.5.1 details the selection of our observational sample, Section 5.5.2 describes the VESPA SFH catalogue, and Section 5.5.3 details our predictions. Finally, in Section 5.6 we discuss our results and avenues for future research, then summarise our conclusions in Section 5.7. We make all of our code for downloading the simulation and observational data, as well as training the CNNs, available online in the form of Jupyter notebooks.[19] Throughout we assume a Planck 2013 cosmology with the following parameters: $\Omega_m = 0.30$, $\Omega_\Lambda = 0.69$, $\Omega_b = 0.048$, $h = 0.68$, $\sigma_8 = 0.83$ and $n_s = 0.96$.

5.2 Methodology

Supervised machine learning methods use training data to learn the relationship between input *features* and output *predictors*. The trained model can then be used to predict values for unseen data. Our features in this work are galaxy SEDs, and our predictors are SFHs. We describe the SFHs as a piece-wise constant curve in bins logarithmically spaced in look-back time:

$$\begin{aligned} 0 < t\,/\,\mathrm{Myr} &< 32 \\ 32 < t\,/\,\mathrm{Myr} &< 68 \\ 68 < t\,/\,\mathrm{Myr} &< 147 \\ 147 < t\,/\,\mathrm{Myr} &< 316 \\ 316 < t\,/\,\mathrm{Myr} &< 681 \\ 0.681 < t\,/\,\mathrm{Gyr} &< 1.47 \\ 1.47 < t\,/\,\mathrm{Gyr} &< 3.16 \\ 3.16 < t\,/\,\mathrm{Gyr} &< 12.46 \ , \end{aligned} \tag{5.1}$$

where t is the lookback time from $z = 0.1$. This choice ensures that the epochs of recent star formation, which leave more significant imprints on the spectrum, are sampled more finely, while older stellar populations that evolve more slowly are grouped in wider bins. The final bin is defined even wider by construction; we tested using higher resolution bins for older populations and found that the machine could not accurately distinguish

[19] https://github.com/christopherlovell/learning_sfhs

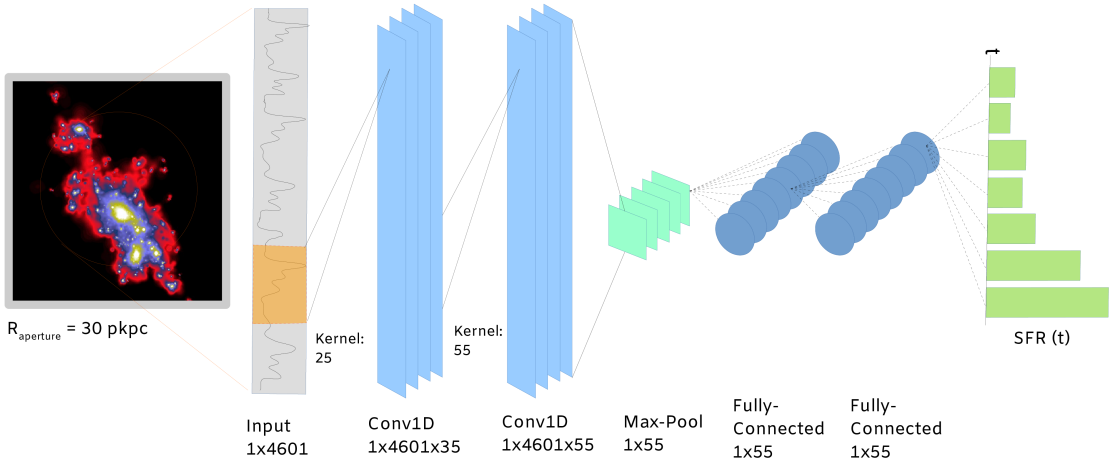

Figure 5.1: The CNN architecture, described in detail in Section 5.2.1.1.

between different aged populations above ~ 3 Gyr.

Before training any of our machine learning methods we first split the data in to training (72%), validation (8%) and test (20%) sets. We take care to perform any optimisation, be that normalisation of the features or hyperparameter optimisation, solely on the training (+ validation) data.

5.2.1 Machine Learning Methods

We implement two different learning algorithms: Extremely Randomised Trees (ERT) and Convolutional Neural Networks (CNN). Using two different methods provides an additional means of evaluating the performance through comparison.

5.2.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs)[20] are growing in popularity in many areas of Astronomy, typically as a means of analysing 2D image data (*e.g.* Tuccillo et al. 2017; Petrillo et al. 2017), and have been shown to perform remarkably well, with prediction accuracies in classification tasks approaching human level (Flamary, 2016; Fabbro et al., 2018).

Our CNN architecture was inspired by the work of Fabbro et al. (2018), who use the python

[20]For further background see Gu et al. (2018); Kiranyaz et al. (2019); Fan et al. (2019)

version of Keras (Chollet et al., 2015) to apply the technique to 1D stellar spectra. We make a number of modifications, as well as a systematic hyperparameter search given our training features. The basic structure (shown in Figure 5.1) uses two convolutional layers, the first applied directly to the one dimensional input spectral features, the latter applied to the outputs of the first layer. The convolution operation essentially shares information between neighbouring pixels, allowing the network to identify spatial correlations in feature space, such as gradients and emission / absorption lines; tiered convolutional layers allow the model to learn higher order relationships. The output of the second convolutional layer is then fed in to a max-pooling layer, which takes the maximum from each feature map generated from the convolutional layers, significantly reducing the dimensionality (from $1 \times 4601 \times 55$ to 1×55); this leads to faster training and reduced overfitting. Finally, the output of the pooling layer is fed in to a traditional fully-connected neural network, where each neuron in a given layer is connected to every neuron in the subsequent layer. We tested different configurations, from shallow and wide (few layers, many neurons in each layer) to deep and narrow (many layers, few neurons in each layer), and settled on the former. The convolution and pooling layers together can be thought of as the *feature extraction* part of the network, and the fully-connected layers perform regression on these features.

The network weights are initially set randomly, then updated through iterations of forward and back propagation utilising the Adam optimizer (Kingma & Ba, 2014). We minimise Symmetric Mean Absolute Percentage Error (SMAPE; see Section 5.2.2) as the target loss function. The network is trained in epochs; during each epoch the training data are fed in batches (the batch size being a free parameter), and once all training galaxies have been used the trained model is evaluated on the validation set. This gives a validation score, that is used to decide when the training has converged, and to prevent overfitting. During training we monitor the validation loss after each epoch and reduce the learning rate if it has plateaued, or stop training altogether if the improvement is below some threshold after a given number of epochs (early stopping), to prevent overfitting.

Optimising the network architecture is notoriously difficult due to the flexibility available in the network configuration. However, once the general architecture has been decided, there are further optimisations that can be made to higher level hyperparameters that

can lead to significant improvements. We use HYPERAS[21] to optimise a subset of these parameters: the number of filters and size of the kernel in each convolutional layer, and the number of neurons in the fully connected layers. Hyperas utilises Tree-structured Parzen Estimators (TPE), which, after an initial random search, sequentially approximates the performance of hyperparameters based on previous measurements, building a likelihood based model (Bergstra et al., 2011).

5.2.1.2 Extremely Randomised Trees

Ensemble decision tree algorithms aggregate the results of multiple trained decision trees in order to produce a single prediction, and can be applied to both classification or regression tasks. Since decision trees are computationally inexpensive to train, the training of ensembles does not lead to a significant performance penalty, and can be simply parallelised. Extremely Randomised Trees (ERT; Geurts et al., 2006) is one such ensemble approach that has been successfully used in a wide range of Astronomy domains (*e.g.* Kamdar et al. 2016; Cohn 2018). It is similar to the popular Random Forest (RF): during training of a RF, a subset of K features is randomly chosen during each split, which reduces the correlation between trees where there are features with a strong correlation with the predictors. ERT also perform this same feature space sampling, but add a further level of randomness by making non-deterministic split choices

We use the implementation of ERT provided in *scikit-learn* (Pedregosa et al., 2011), with grid search cross validation to optimise the following hyperparameters: minimum samples in a split, minimum samples in a leaf, and maximum nodes in a leaf. This optimisation is done solely on the training set during each training procedure. For ERT, the full training set (training + validation) is used during training and optimisation.

5.2.2 Loss Functions

During model training and evaluation, the fit is assessed through a particular *loss function*. Typical loss functions include the mean absolute percentage error (MAPE) and the mean squared error (MSE), with the mean taken over all of the output predictors. Both of these loss functions are inappropriate when applied to star formation histories sampled

[21] https://github.com/maxpumperla/hyperas

from a reasonably wide range of final stellar masses. For example, the MSE leads to large penalties for histories with high SFH normalisation, whilst lower mass galaxies with a lower SFH normalisation are not penalised to the same degree despite similar percentage errors in their predictions. On the other hand, it is not possible to calculate percentage errors for zero valued bins.

We would ideally like a loss function that acts as a percentage error, in order not to penalise high mass galaxies, but returns reasonable results for zero valued bins. We use a variation of Symmetric Mean Absolute Percentage Error (SMAPE),

$$\text{SMAPE} = \left[2 \times \frac{\Sigma_b \, |\mathrm{Y}_b^{\mathrm{true}} - \mathrm{Y}_b^{\mathrm{pred}}|}{\Sigma_b \, (\mathrm{Y}_b^{\mathrm{true}} + \mathrm{Y}_b^{\mathrm{pred}})} \right] \times 100\% \ ,$$

where Y_b is the star formation rate in bin b. The value of SMAPE is bounded between $0\% < \text{SMAPE} < 200\%$, but acts as a true percentage error in the low error regime. This point statistic can be used as both a reasonably unbiased loss function within the CNN, and as an evaluation of the fit.

5.2.3 Cosmological Simulations

We use two cosmological hydrodynamic simulations, EAGLE (Schaye et al., 2014; Crain et al., 2015) and Illustris[22] (Vogelsberger et al., 2014; Genel et al., 2014), which have both been run on large comoving volumes, tens of megaparsecs on a side, producing tens of thousands of galaxies at $z = 0$. EAGLE[23] uses a modified version of the Smoothed Particle Hydrodynamics (SPH) code GADGET 3 (Springel et al., 2005), whereas Illustris uses the moving-mesh code AREPO (Springel, 2010b). The typical gas element mass in each simulation is $\sim 10^6 \, M_\odot$; below this mass scale physical processes cannot be modelled self consistently, so subgrid prescriptions are used to handle processes such as radiative cooling, star formation, stellar evolution, star formation feedback, black hole seeding, and AGN feedback. Each hydrodynamic solver handles shocks and instabilities differently, but on the whole the choice of solver does not have a large effect on global galaxy properties; it is in the subgrid models that significant differences between the simulations are most

[22]Galaxy and particle information for Illustris were obtained from the online API, http://www.illustris-project.org/data/

[23]Galaxy and particle information for EAGLE were obtained from the public database, http://icc.dur.ac.uk/Eagle/database.php (McAlpine et al., 2016; The EAGLE team, 2017)

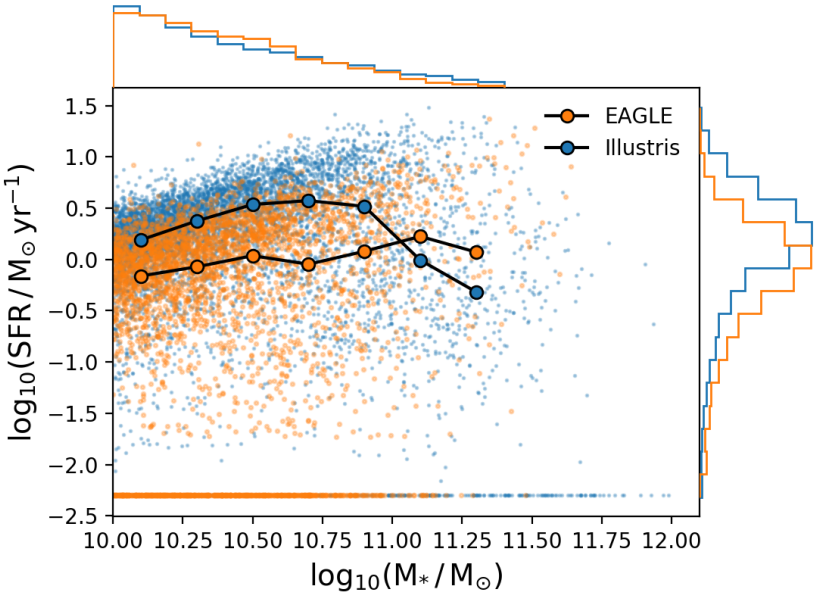

Figure 5.2: The M_* - SFR relation, or star-forming sequence, at $z = 0.1$ for the selected Illustris and EAGLE galaxies. The scatter shows individual objects, and the median relation with 1σ spread is over-plotted. SFR is calculated using the integrated mass of stars formed in the last 100 Myr within a 30 pkpc aperture. Galaxies with zero recent SFR are plotted at $10^{-2.3}$ $\mathrm{M}_\odot\,\mathrm{yr}^{-1}$ for clarity. The histograms at the top and right of the plot show the normalised number counts as a function of stellar mass and SFR, respectively. EAGLE and Illustris predict contrasting behaviour on this parameter plane.

apparent (Somerville & Davé, 2015).

By using two different simulations we are able to evaluate how our algorithms generalize, by training them on a single simulation then testing its performance on another. We can then assess whether we are learning the *intrinsic* relationship between galaxy SEDs and their SFHs, rather than learning about the relationship in a particular simulation.

Both EAGLE and Illustris have been shown to agree reasonably well with observed stellar mass and star formation rate distribution functions at low redshift, though there are still discrepancies both between the simulations and with the observations. For example, EAGLE fits the low mass end of the Galaxy Stellar Mass Function (GSMF), but underestimates the normalisation at intermediate masses around the knee of the GSMF ($\sim 5 \times 10^{10}\,M_\odot$), whereas Illustris overestimates both the low mass and high mass number densities, but shows good agreement around the knee (Schaye et al., 2014; Genel et al., 2014). Even greater discrepancies between the simulations can be seen in the distribution of specific Star Formation Rate (sSFR = SFR / M_*) as a function of stellar mass, which in EAGLE shows a relatively flat relation up to $M_* / M_\odot \sim 10^{10}$, which then falls by ~ 0.8

dex; this agrees with the observations, but the normalisation is ~ 0.3 dex lower at all but the highest stellar masses (Schaye et al., 2014). In contrast, Illustris remains flat out to $M_* / M_\odot \sim 10^{11}$ (Sparre et al., 2015); Illustris galaxies with Milky Way-like masses exhibit higher SFRs compared to EAGLE.

Such differences are to be expected due to the complexity of physical processes to be modelled at a large range of scales, and their resolution is a key goal of research in the field. However, confusingly, the photometric colour distributions in both simulations have been shown to be in relatively good agreement with observations at low redshift over the same mass range (Trayford et al., 2015; Vogelsberger et al., 2014). This inconsistency, between the intrinsic physical properties and the predicted photometric distributions, is due to differences in the choice of SED modelling assumptions, particularly the magnitude of the dust correction.

Both simulations assume a Chabrier IMF, but adopt different cosmological parameters; Illustris assumes WMAP9 (Hinshaw et al., 2013), EAGLE Planck13 (Planck Collaboration et al., 2014), however these differences are expected to have negligible impact on the resulting galaxy distribution functions.

5.2.3.1 Measurement Aperture

A significant proportion of the stars in massive galaxies are located within an extended halo surrounding the central stellar concentration. These stars tend to be older, are often accreted from other systems through interactions, and therefore have a different SFH from those in the centre, which leads to spatial gradients in physical and observed stellar properties. Both the integrated luminosity and the colour of a galaxy are therefore sensitive to the measurement aperture, and in order to facilitate comparison with observations similar apertures should be used when generating synthetic SEDs. Unfortunately, this relies on the simulations having realistic spatially resolved star formation histories, which has not been extensively tested, and is also subject to resolution issues for small apertures. We use a spherical 30 kpc aperture centred on the gravitational potential minimum, which has been shown to yield similar masses to a Petrosian aperture typically used in photometric observational studies (Schaye et al., 2014). All quoted stellar properties (M_*, SFR, SFH, etc.) are taken from the star particles within this aperture, and synthetic

spectra are generated using only these star particles (see Section 5.2.4); this must be taken in to account when comparing to observational studies (see Section 5.5.1).

5.2.3.2 Galaxy Selection

We select all galaxies from each simulation at $z = 0.1$ with stellar masses $\mathrm{M}_* \,/\, \mathrm{M}_\odot > 10^{10}$, which gives 3687 and 6473 galaxies for EAGLE and Illustris, respectively. The large offset is an unfortunate result of the difference in the GSMF normalisation between the simulations at the high mass end. Figure 5.2 shows the distribution of our selections on the $\mathrm{M}_* - \mathrm{SFR}$ plane. The normalised histogram at the top shows the distribution of stellar masses; the steepness of the GSMF in both simulations means that their are many more low mass galaxies than high. Since these low mass galaxies dominate our training sample, we expect to see a degree of overfitting to such galaxies with respect to their less numerous high mass counterparts. We explore this in more detail in Section 5.3.

Illustris shows a steeper star-forming sequence relation than EAGLE and a higher normalisation between $10 < \log_{10}(\mathrm{M}_* \,/\, \mathrm{M}_\odot) < 11$, but above this Illustris galaxies have lower SFRs. Such significant differences in training and test data present a unique challenge for machine learning methods, where the accuracy on unseen data is usually poor, and as such represents a robust test of our method.

5.2.4 Synthetic Spectra

The composite spectrum of a galaxy in each simulation is dependent upon the physical properties and spatial distribution of the stars, gas and black holes. We ignore the AGN contribution, which we do not expect to have a great effect on the optical emission.

The pipeline for generating spectra detailed in this section is contained within the SPECTACLE module, available at https://github.com/christopherlovell/spectacle.

5.2.4.1 Intrinsic Spectra

We generate intrinsic spectra by treating each star particle as a simple stellar population (SSP). We then generate an SED for each SSP using the Python implementation of the Flexible Stellar Population Synthesis (FSPS) code (Conroy et al., 2009; Conroy & Gunn, 2010; Foreman-Mackey et al., 2014). The SED of each SSP is dependent on its age and

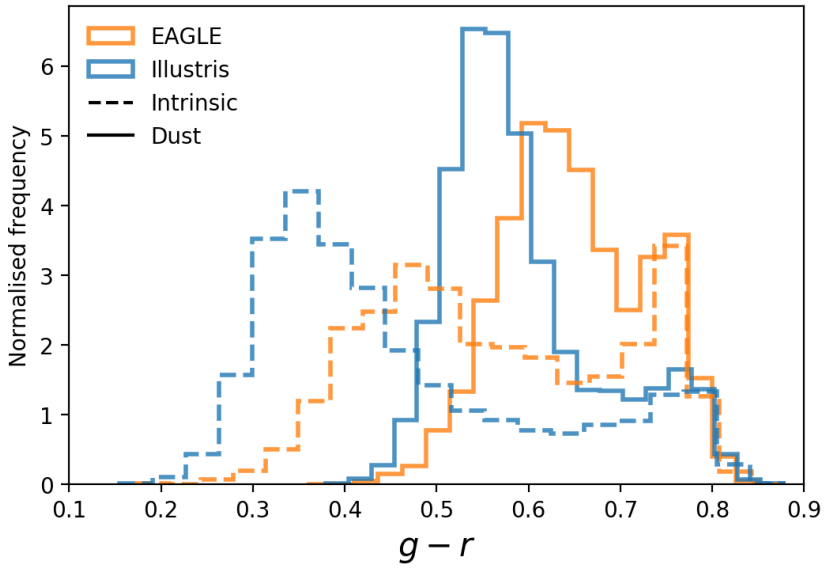

Figure 5.3: $g - r$ colour distribution for the EAGLE and Illustris simulation selections. Dashed lines show the intrinsic distributions (including the nebular contribution); solid lines show the dust-attenuated distributions. The dust model leads to a significant reddening of the blue population in both simulations.

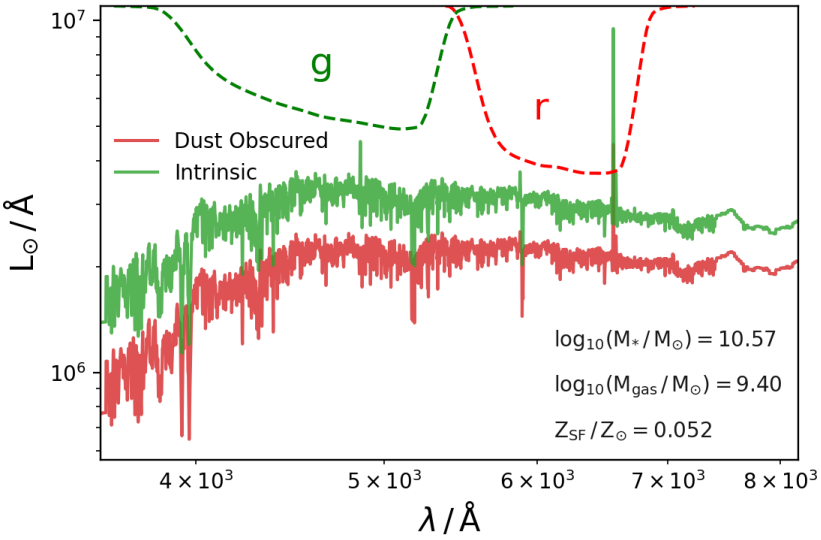

Figure 5.4: Intrinsic (green) and dust-obscured (red) spectrum for an example galaxy from the Illustris simulation. The g and r filter curve responses are shown at the top of the plot.

metallicity, normalised by its initial stellar mass. Each stellar particle in the simulations is approximately two orders of magnitude more massive than typical star forming regions; a single young star particle can therefore significantly affect the predicted colours of a galaxy. In order to mitigate this artificial Poisson scatter we resample the recent star formation using a similar technique to that used in Trayford et al. (2015). We take each star particle younger than 100 Myr and split it into ten thousand new particles with ages sampled uniformly within this interval, and the mass of the original particle equally distributed between the resampled particles.

Young stellar populations ionise their surrounding gas, leading to nebular line and continuum emission. This emission can dominate photometric fluxes, as well as being responsible for the majority of optical emission lines (Anders & Fritze-v. Alvensleben, 2003; Reines et al., 2010; Wilkins et al., 2013b). Byler et al. (2017) use the photoionization code CLOUDY to model the expected nebular emission from young FSPS SSPs self-consistently; these templates are provided in python-FSPS. They assume a covering fraction of unity for stellar populations with age $t < t_{\rm esc}$, where $t_{\rm esc} = 10^7$ years.

We define the 'intrinsic' emission as including the nebular contribution. Figure 5.3 shows the intrinsic $g - r$ colour distribution for EAGLE and Illustris. Figure 5.4 shows the example intrinsic emission for an Illustris galaxy; strong nebular line emission and absorption are clearly visible.

5.2.4.2 Dust Attenuated Spectra

We use information from the models on the mass and metallicity of star forming gas to provide a self-consistent, physically motivated prescription for the dust attenuation. Our model assumes a simple screen, ignoring the dust distribution geometry. The transmission T at a wavelength λ for a particle of age t is given by

$$T(\lambda, t) = \exp\left[-\tau(t)\left(\frac{\lambda}{\lambda_\nu}\right)^{-1}\right] \quad , \tag{5.2}$$

where τ is the optical depth at wavelength λ_ν. The optical depth is dependent on the age of the stellar particle; all particles are subject to a constant screen due to dust in the ISM, but young particles, which still reside within their birth clouds, are subject to a further

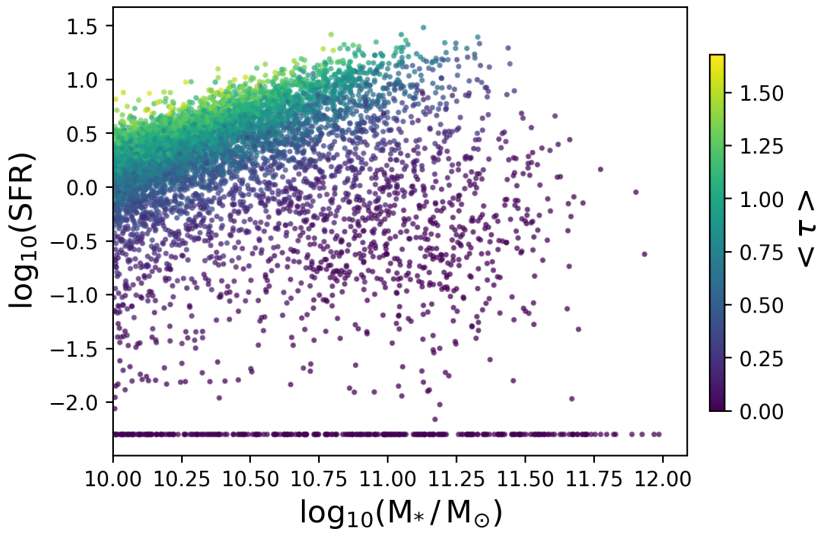

Figure 5.5: The star-forming sequence for the Illustris sample, coloured by the average attenuation over the whole galaxy ($< \tau >= -\log(F_\lambda^{\rm dust} \,/\, F_\lambda^{\rm int})[\lambda = 5500$ Å]). Gas-rich, star-forming galaxies experience greater attenuation than gas-poor galaxies at the same stellar mass.

transient attenuation component,

$$t \leq t_{\rm disp} \;:\; \tau = \gamma\,\tau_{\rm cloud} + \gamma\,\tau_{\rm ISM}$$
$$t \geqslant t_{\rm disp} \;:\; \tau = \gamma\,\tau_{\rm ISM} \quad .$$

Both $\tau_{\rm ISM}$ and $\tau_{\rm cloud}$ can be fixed constants ($\gamma = 1$), or linked to other properties of the galaxy. We link the optical depth to the metallicity and mass of cold, star forming gas:

$$\gamma = \frac{Z_{\rm SF}}{Z_{\rm Z14}} \left(\frac{M_{\rm SF}}{M_*} \frac{1}{\beta} \right) \quad , \tag{5.3}$$

where $Z_{\rm SF}$ is the mass-weighted star forming gas phase metallicity, and the mass dependence is encapsulated in the ratio of $M_{\rm SF}$, the total mass of star forming gas, to M_*, the stellar mass. These are both normalised to the respective Milky Way values: $Z_{\rm Z14} = 0.035$ [24], and $\beta = 0.1$. We use $\tau_{\rm cloud} = 0.67$, $\tau_{\rm ISM} = 0.33$, $t_{\rm disp} = 10\,$Myr and $\lambda_\nu = 5500\,$Å, as used in both EAGLE and Illustris studies (Trayford et al., 2015; Genel et al., 2014). This approach produces a physically motivated attenuation, where gas rich spirals are subject to higher attenuation than gas poor ellipticals with identical stellar

[24] This is taken from the $M_* - Z$ relation expression in Zahid et al. (2014) evaluated at the Milky Way stellar mass, and converted to relative solar metallicities assuming $12+\log_{10}({\rm O/H})_\odot = 8.69$ (Allende Prieto et al., 2001).

mass. This can be seen in Figure 5.5, which shows the star-forming sequence for the Illustris selection, coloured by the mean attenuation.

Figure 5.4 shows the dust-obscured spectrum for an example Illustris galaxy. The high relative gas mass and star-forming gas phase metallicity leads to significant attenuation. Figure 5.3 shows the distribution of $g - r$ colour for the dust attenuated spectra. Dust leads to a reddening of the blue population, shifting the peak by $\Delta(g - r) \sim +0.2$ in both simulations, but the location and normalisation of the red population in both cases is generally unaffected; this is expected since these intrinsically red systems are generally gas poor, and experience lower attenuation.

5.2.4.3 Artificial Noise

In order to further increase the realism of our synthetic spectra we add artificial noise at a given signal to noise (SN) level. We use a fiducial value of SN = 50, and test the effect of increased SN on our predictions in Section 5.3.1.

For each spectrum we can take multiple realisations of the noise. This can be useful in two ways: it can increase our training set, and it can prevent the model from overfitting to a single noisy realisation by providing multiple noise-added spectra for a given SFH. We explore the effect of using multiple noisy realisations on our model training in Section 5.3.1.

5.2.4.4 Wavelength Grid

We restrict the wavelength coverage to that approximately covered by the SDSS DR7 release (see Section 5.5.1), and resample (flux preserving; Carnall, 2017) on a fixed logarithmically-sampled wavelength grid. This gives a final fixed input wavelength grid, $3572 \leqslant \lambda \,/\, \text{Å} \leqslant 8173$, with resolution $\lambda \,/\, \Delta\lambda = 5570$ $(\lambda = 4500\,\text{Å})$.

5.3 Results

We first train both Extremely Randomised Trees (ERT) and Convolutional Neural Network (CNN) models on our EAGLE and Illustris training samples (80% of the data). All plots in this section show predictions when applied to galaxies in the respective test sets (20% of the data).

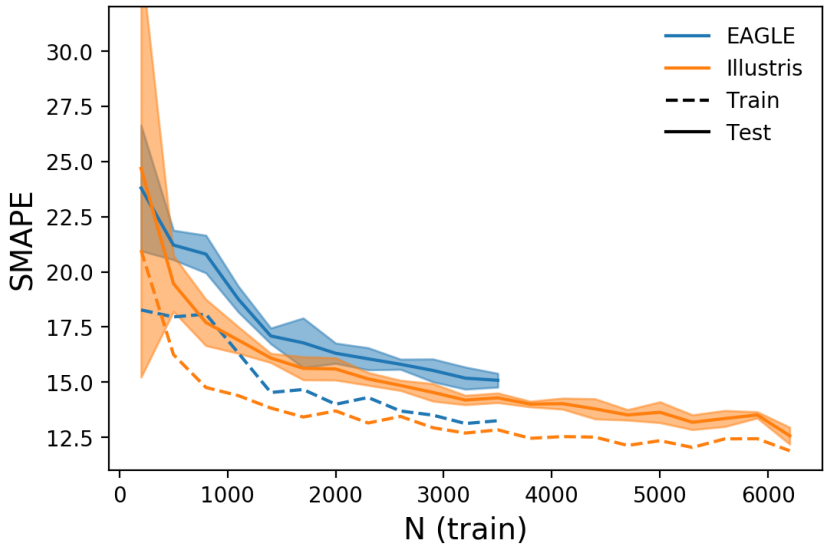

Figure 5.6: Learning curves, showing the SMAPE as a function of input training data size, from CNN trained on dust attenuated spectra from both Illustris and EAGLE. Multiple samples without replacement are drawn from the full training set, and the median SMAPE on the training and test sets are shown as the dashed and solid lines, respectively. The shaded region showing the 1σ spread in the test SMAPE.

5.3.1 Training & Testing

5.3.1.1 Learning Curves

Learning curves show the improvement in test score as a function of training set size, which provides information on the convergence of the model. Decreasing scores suggest that a larger training set would lead to a better fit, whereas a plateau suggests that the training has converged and no further improvement can be obtained from additional training data. A large gap between the training error and the test error would indicate overfitting, or poor generalisation properties. Figure 5.6 shows learning curves for dust attenuated spectra from Illustris and EAGLE. We perform 6-fold cross validation to estimate the scores and present their median. The EAGLE learning curve is still falling at 3500 samples, which suggests that the model is yet to converge. The Illustris learning curve, in comparison, appears to have plateaued at $\sim$ 5500 samples, though a larger training set is needed to confirm this. As a result, we concentrate on the converged Illustris model for the time being (we will return to the EAGLE training set later, both in conjunction with the Illustris training data, and as an independent test set for the Illustris trained model). The gap between the training and test errors in both EAGLE

and Illustris is small, which suggest negligible overfitting. The EAGLE model has slightly higher SMAPE at fixed N than Illustris, but it is unclear what specific differences in the simulation modelling lead to this; a possible explanation is the higher gas-phase metallicity in EAGLE at fixed stellar mass compared to Illustris, which will contribute to greater dust attenuation, obscuring the underlying relationship between the SFH and the spectra more in EAGLE than Illustris.

5.3.1.2 Method comparison

Returning to the Illustris data in isolation, the top panel of Figure 5.7 shows the distribution of SMAPE scores, for both ERT and CNN and for dust obscured and intrinsic spectra, evaluated over the entire Illustris test set. The median SMAPE for the CNN is significantly lower than for ERT for both intrinsic and dust obscured features. This is due to the CNN's ability to share local information between neighbouring pixels, whereas ERT treats each pixel as an isolated feature. We also find that the median SMAPE for dust obscured spectra with ERT is significantly higher than that for intrinsic spectra, however for the CNN this difference is negligible. Dust introduces additional degeneracies between the spectral features and the underlying SFH, so it interesting that the CNN is capable of overcoming these.

We choose to focus on the CNN performance in the rest of the paper.

5.3.1.3 Model Results with Noise

As mentioned in Section 5.2.4.3 we add noise to our simulated spectra with a fiducial value of SN=50. The middle panel of Figure 5.7 shows the SMAPE distribution for a model trained with this added noise, and as expected the noise leads to an increase in the median SMAPE of 2%. However, we can re-sample the noise for each synthetic spectrum multiple times. Using a multi-resampled training set leads to a reduction in the SMAPE; we tested different numbers of resamples, and found that the improvement in SMAPE plateaus at 4. The SMAPE distribution using this 4 times resampled feature set is shown in the middle panel of Figure 5.7; the median SMAPE is much lower than for the single noise-realisation feature set (10.9%). This suggests that the negative effect of the noise, that obscures the relationship between the spectra and the SFH, is overcome by the positive impact of the larger, more generalisable training set.

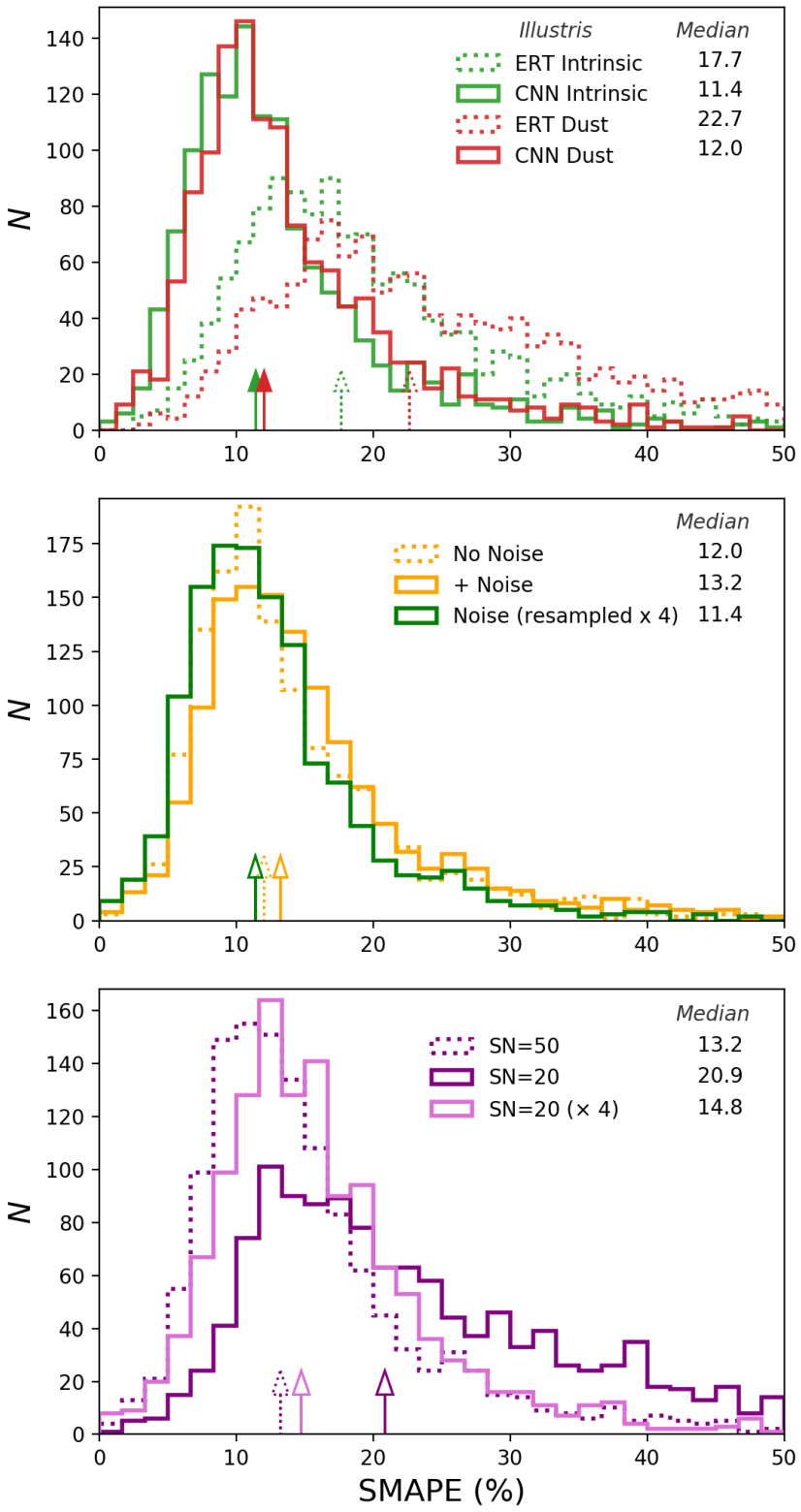

Figure 5.7: SMAPE distributions for the Illustris simulation, with different learning algorithms and spectral modelling. The median of each distribution is shown by the arrows, and quoted in the legend. *Top:* ERT (dashed) and CNN (solid) models trained on intrinsic (green) and dust-obscured (red) spectra. *Middle:* CNN model trained on dust-obscured spectra (dashed), with added noise (solid, yellow), and with noise resampled × 4 (solid, green). *Bottom:* CNN model trained on dust-obscured spectra with added noise at SN=50 (dashed, purple), SN=20 (solid, purple), and with noise resampled × 4 at SN=20 (solid, pink).

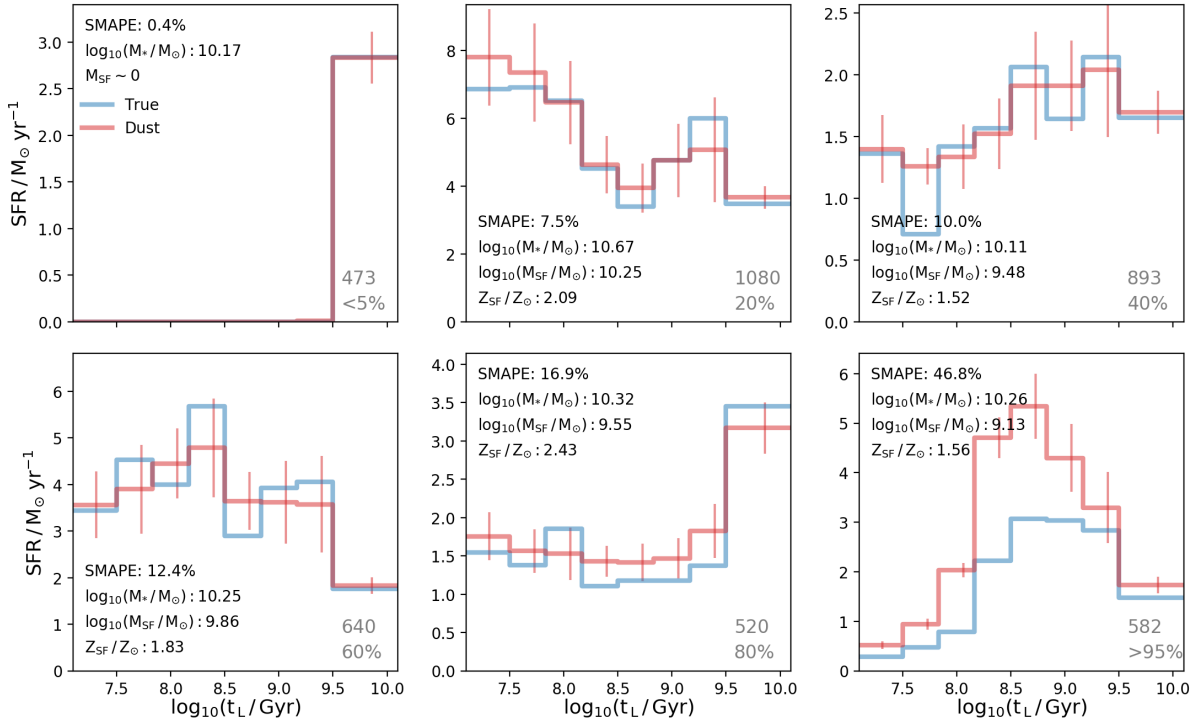

Figure 5.8: Six example SFHs from the Illustris test set (blue), alongside fits to the dust-obscured spectra (red). The examples are selected with a range of SMAPE scores, 0.8-55.6%, from top left to bottom right. Errors are a combination of observational and modelling errors, see Section 5.4. Each panel shows the galaxy index and the approximate SMAPE score percentile in the bottom right, as well as the $z = 0$ stellar mass, star-forming gas mass and star-forming gas metallicity.

We expect the prediction accuracy to decrease as the noise level is increased. To test this, we used a SN = 20, shown in the bottom panel of Figure 5.7. This leads to an increase in the SMAPE of 2.9% compared to the fiducial SN = 50. However, as in the lower noise case, resampling the noise 4 × leads to an improvement of 1.9% in the median SMAPE over the single-realisation model. We quote results using the SN=50, 4 × resampled spectra in the rest of this section, unless otherwise noted.

5.3.1.4 Example Fits

In order to illustrate the SFH fits we show six examples from the Illustris test set in Figure 5.8. We show predictions for a range of SMAPE scores as evaluated on the dust attenuated SEDs. The top left panel shows one of the best fits, the next four panels show fits around the 20th, 40th, 60th and 80th percentiles of the SMAPE distribution, and finally the bottom right panel shows one of the worst fits. The errors on the fit in each bin are taken from the observational and model errors combined in quadrature (see Section 5.4).

5.3.1.5 Parameter Correlations

As mentioned in Section 5.2.3.2, we preferentially select low mass galaxies due to the steepness of the GSMF. It is therefore important to investigate any correlation of the quality of fit with stellar mass, to evaluate any overfitting to low mass galaxies. The top panel of Figure 5.9 shows the distribution of Illustris test galaxies on the stellar mass - SFR plane, coloured by SMAPE on the predicted histories from the dust-attenuated model. In order to quantify any trend of SMAPE with our galaxy parameters we calculate the Pearson's correlation coefficient,

$$\rho = \frac{\mathrm{cov}(P, SMAPE)}{\sigma_P \sigma_{SMAPE}} \quad ,$$

where P is the given parameter, cov is the covariance between the parameter and SMAPE, and σ is the standard deviation of the respective quantity. There is no significant correlation between stellar mass and SMAPE ($\rho = -0.14$), nor between specific-SFR and SMAPE ($\rho = 0.11$).

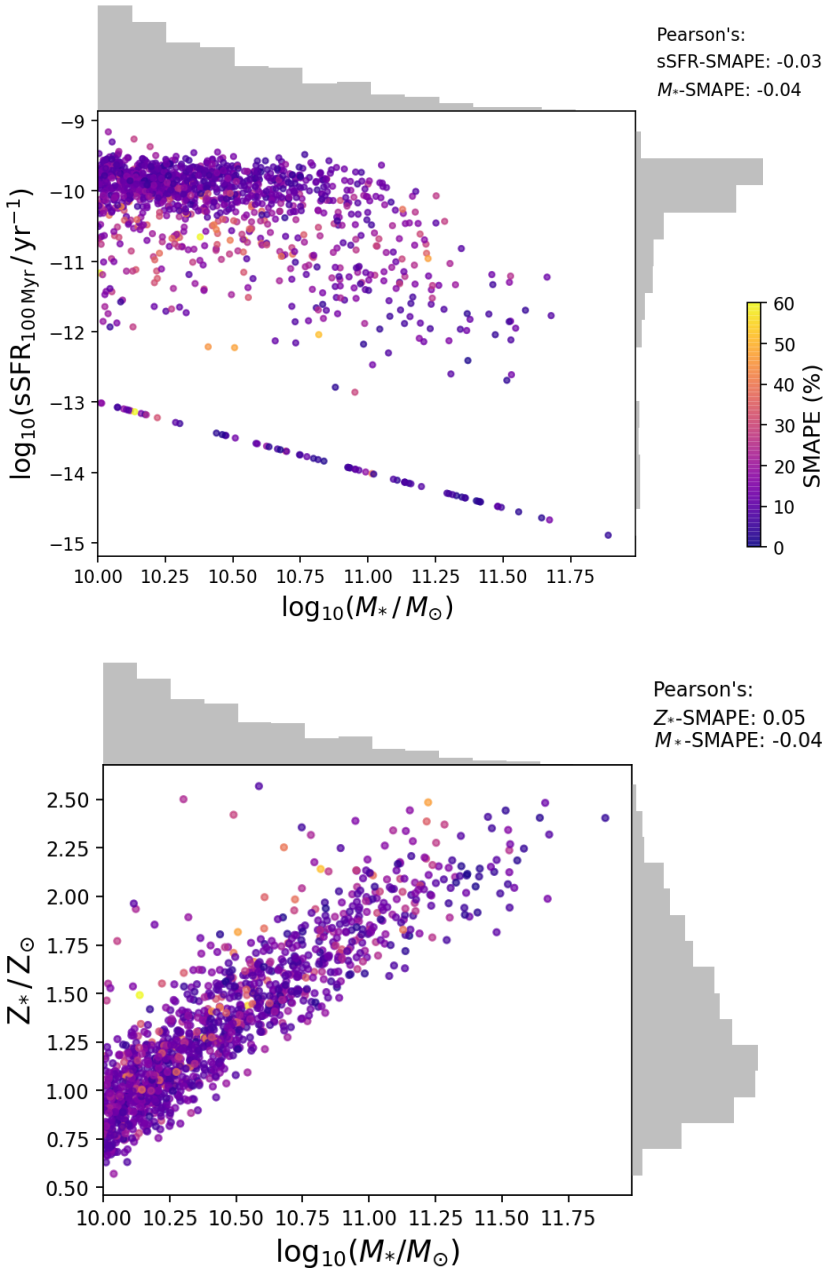

Figure 5.9: Parameter correlations with SMAPE for the predictions on the Illustris test set, using the intrinsic spectra. The pearson's correlation coefficient between each parameter and SMAPE is shown in the top right. The grey histograms above and to the right of each axis show the distribution of the given parameter. *Top:* stellar mass - SFR relation. SFR is calculated as the integrated mass in stars formed in the last 100 Myr. *Bottom:* stellar mass - stellar metallicity relation.

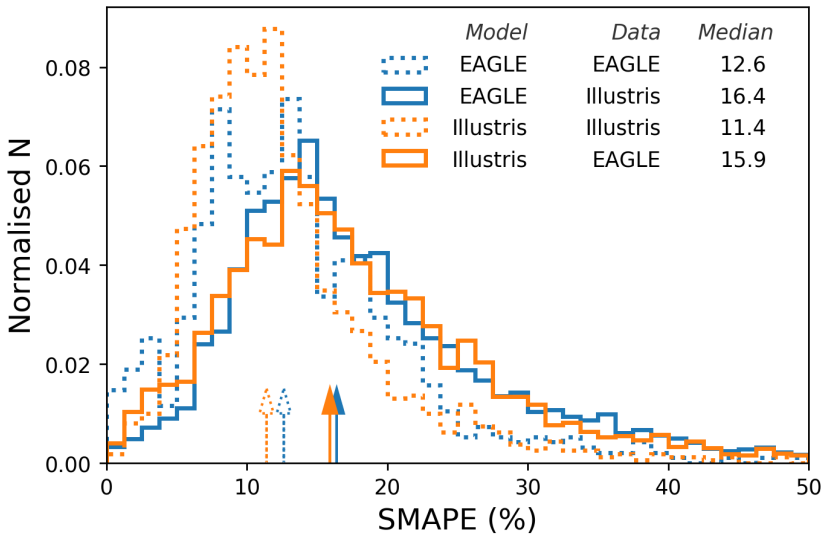

Figure 5.10: The normalised SMAPE distribution for the inter-sim (solid) and within-sim (dashed) test sets, for dust-attenuated spectra. The median of the distribution is shown by the arrow on the x-axis, and quoted in the legend. Despite being trained on very different data, the SMAPE is low in both inter-sim cases.

The well known age-metallicity degeneracy in the optical can also obscure the underlying SFH (Worthey, 1994). The bottom panel of Figure 5.9 shows the stellar mass - metallicity distribution, for the Illustris test galaxies, coloured by SMAPE on the intrinsic model. There is no significant correlation between stellar metallicity and SMAPE ($\rho = -0.05$). This may be due to the relatively low resolution of the SFHs, reducing the confusion between bins.

5.3.2 Testing Across Simulations

Further uncertainty is introduced by our choice of modelling assumptions, such as the training simulation, SPS model, intrinsic SED pipeline and dust model. Of these we expect the choice of training simulation to lead to the greatest bias. To estimate the uncertainty introduced we test a model trained assuming some simulation training data on another model trained assuming different simulation training data. This procedure demonstrates how well each model generalises.

Figure 5.10 shows the SMAPE error when our CNN is trained and tested on different simulations, using dust-obscured spectra with 4 $\times$ resampled noise. Since the latter testing simulation is not included in any of the training, the full galaxy sample can be used for testing; we plot the normalised distributions to aid comparison. For models

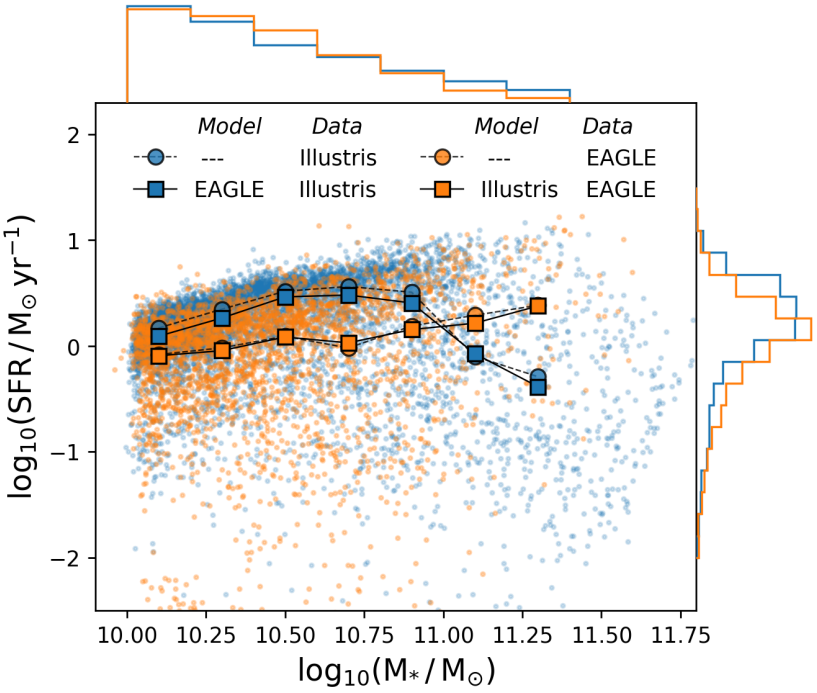

Figure 5.11: The predicted star-forming sequence for the intersim results. We estimate the present day SFR from the normalisation in the latest SFH bin, corresponding to a timescale of approximately 30 Myr, and the total mass from the SFH combined with an age-dependent recycling fraction. Each model prediction, shown with the square points and solid lines, recovers the original star-forming sequence, shown by the circular points and dashed lines, despite being trained on SFHs corresponding to a different SFR-M_* relationship.

trained on both EAGLE and Illustris the median SMAPE for the intra-sim results is higher than within-sim. The errors are still reasonably good in all intra-sim cases, despite the significant differences in the simulations used for the training and testing data.

Another way of testing whether the model is overfitting is to plot the predicted distribution of galaxies on the stellar mass-star formation rate plane. We have already seen in Figure 5.2 that both simulations exhibit very different behaviour in this space, and might expect a model that has overfit to a particular simulation to recover the distribution from its training data. Figure 5.11 shows that this is not the case: each model recovers the star-forming sequence of the new input data.

Whilst these integrated and point-in-time properties are recovered accurately, the shape of the SFH, and the distribution of SFHs, may still be incorrectly predicted. To test this we show in Figure 5.12 the median and $16^{\text{th}} - 84^{\text{th}}$ percentile spread in each bin for the input data and the predictions. The distribution of predicted SFHs is remarkably similar for both simulations throughout cosmic time.

5.4 Error Estimates

Our SFH predictions are subject to two main sources of uncertainty: those from errors in the spectra, which we refer to as *observational* errors, and those from errors in the CNN fit, which we refer to as *modelling* errors. In this section we make estimates for the impact of these two sources of error, and combine them to give a total estimated error in each bin.

5.4.1 Observational Errors

Errors in the observed SED will lead to uncertainty in the predicted histories. The propagated error can be estimated in two ways: sample a number of noisy SEDs, predict the SFHs for each noise-added spectrum, and calculate the covariance matrix of the output, as in Tojeiro et al. (2009), or treat the model as a vector valued function and evaluate the dot product of the Jacobian and the error spectrum, as demonstrated in Fabbro et al. (2018). Errors calculated with both procedures should give similar results since they are essentially evaluating the same input dependence; the former does this through Monte Carlo sampling, whereas the latter explicitly calculates the gradient of the

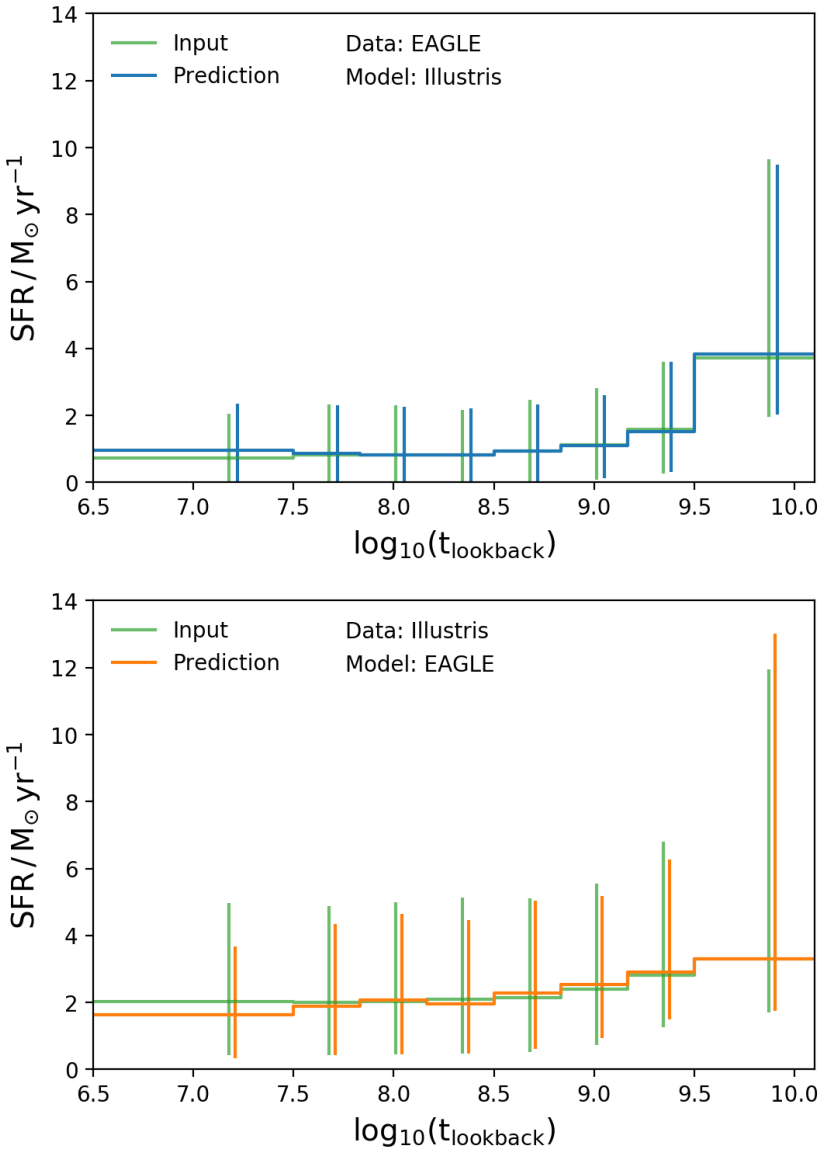

Figure 5.12: The median SFH and $16^{\text{th}} - 84^{\text{th}}$ percentile spread in each bin for the input data (green) and the intersim prediction (orange for the EAGLE mode, blue for the Illustris model). The distribution of predicted SFHs is recovered well in both cases.

predictors with respect to the features.

We implement the former approach, using the noise model described in Section 5.2.4.3. For each spectrum we add N random realisations of each error spectrum to the input spectrum, and propagate each noise-added spectrum through our model to obtain a distribution of predicted histories. From these the covariance matrix can be calculated,

$$C_{ij} = \langle (x_i - \hat{x}_i)(x_j - \hat{x}_j) \rangle \quad ,$$

where x_i is the SFR in bin i for a given realisation, and $\hat{x}_i$ is the mean SFR in that bin for all realisations. The uncertainty in each bin is then $\sigma_i = \sqrt{C_{ii}}$. We can also use C to find the correlation matrix; we describe this in more detail, alongside examples, in Appendix 5.8.1.

Figure 5.13 shows the observational error in each bin as a function of SFR, for intrinsic and dust-obscured spectra. The error is positively correlated with the quantitative value of the SFR. In all but the oldest bin, the errors on dust attenuated spectra are larger than in the intrinsic case. We fit second order polynomials to the $\sigma - \log_{10}(\mathrm{SFR})$ relation for each bin, which allows us to predict the observational error for arbitrary histories (fit parameters are quoted in Appendix 5.8.2).

5.4.2 Modelling Uncertainties

There are a number of free parameters in our model pipeline, from the synthetic SED generation to the parametrisation of the dust model, to the free parameters of the CNN. It is impractical to estimate the uncertainty on each parameter, however we can obtain an estimate of the propagated model uncertainty directly from the scatter of the residuals in predicted SFH. The magnitude of the residual is SFR dependent in all bins; we account for this by dividing by the absolute predicted SFR in the bin to give the *fractional residual*. This single statistic can be used to estimate the model error for each galaxy, bin-by-bin, by multiplying by the predicted SFR.

Figure 5.14 shows the fractional residuals between the predicted and the true SFR in each lookback-age bin as a function of the true SFR within that bin, along with normal fits to the marginalised distributions.

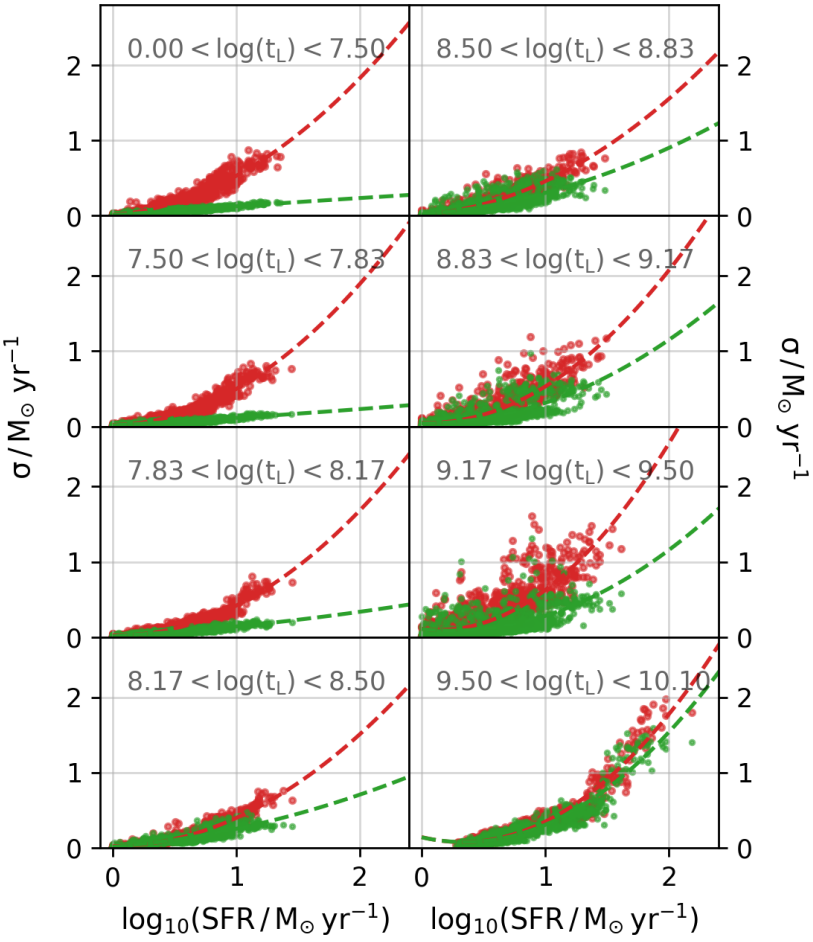

Figure 5.13: Observational errors (1σ) as a function of SFR in each bin, for intrinsic (green) and dust-obscured (red) spectra. Second order polynomial fits are shown as dashed lines. Observational errors are strongly dependent on the quantitative SFR, and are larger for dust-obscured spectra in recent bins.

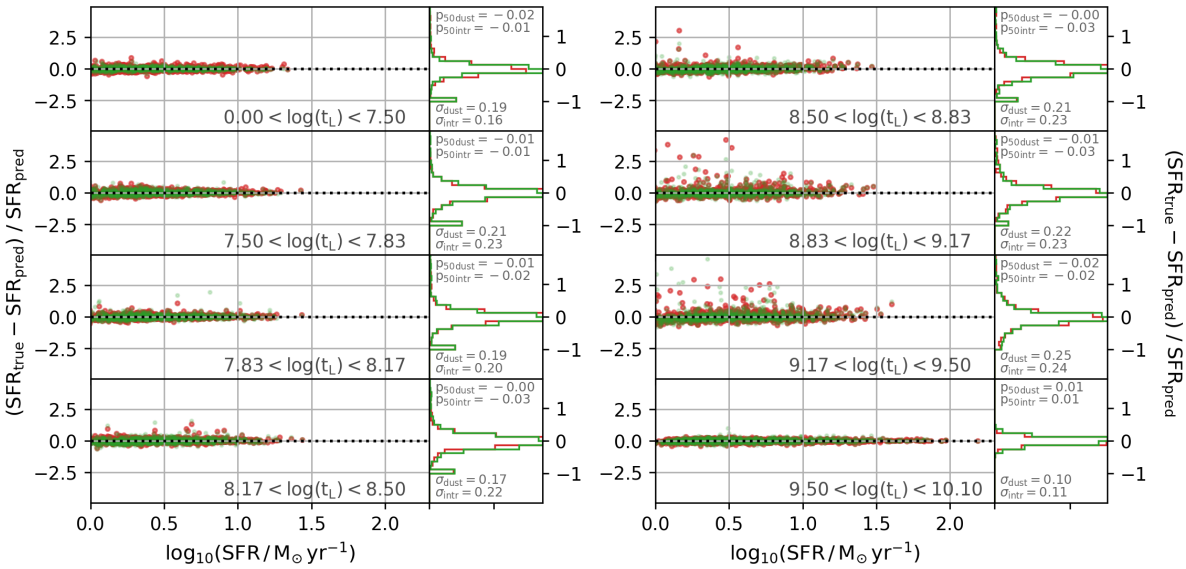

Figure 5.14: Fractional residuals between the true SFH and the predicted SFH for intrinsic (green) and dust attenuated (red) spectra from Illustris. The residuals are plotted as a function of the logarithm of the absolute star formation. The right panels show a one dimensional histogram of the distribuion of fractional residuals, with mean and 1σ spread from a normal fit quoted in each panel.

5.4.3 Total Error

We combine the observational and modelling errors to obtain the *total* error by adding them in quadrature. Since the error is dependent on the quantitative SFH in each bin we do not quote it, but provide fits to the observational error and fractional residual distributions in Appendix 5.8.2. The modelling errors dominate the error budget for all bins, for an observational error SN = 50; we have tested up to SN = 20, and this remains the case. Figure 5.8 shows the total uncertainties calculated using this method, for each example.

5.5 Observational Predictions

We apply the model to the SDSS DR7 Main Galaxy Sample (MGS)[25] (Strauss et al., 2002; Abazajian et al., 2009), which allows us to compare with VESPA (Tojeiro et al., 2007, 2009), an SED fitting code for predicting SFHs that has been applied to this catalogue. VESPA uses similar binned star formation histories to our method, allowing a like-for-like comparison between the two methods. The level of agreement in predicted SFHs, or lack

[25]obtained from the Data Archive Server, das.sdss.org

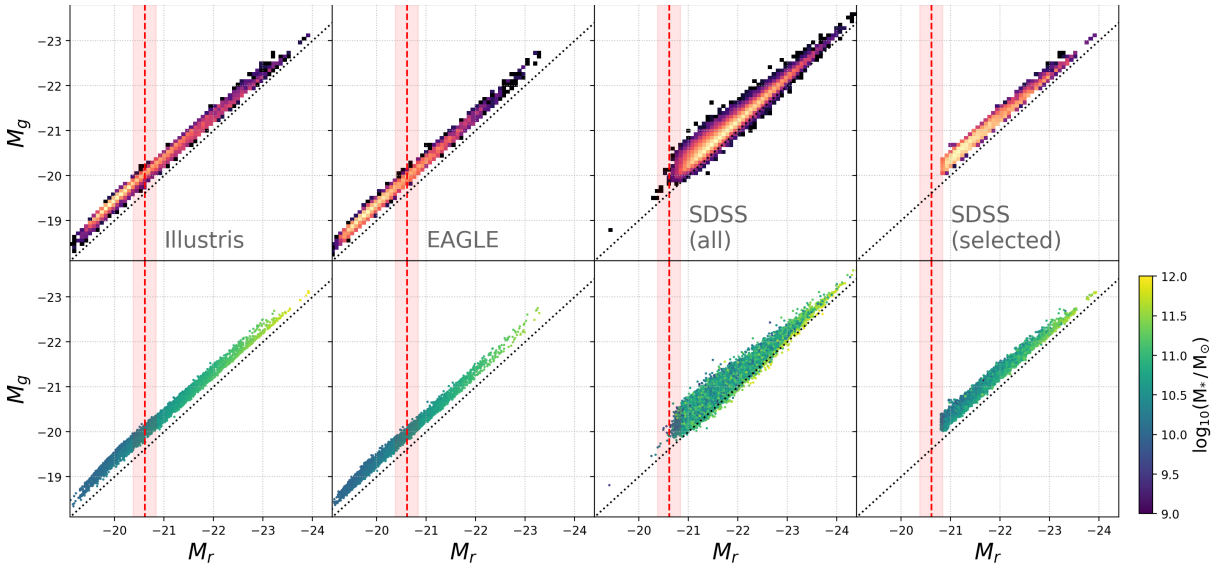

Figure 5.15: g' and r' magnitude distributions in EAGLE, Illustris, all SDSS galaxies, and our final magnitude- and mass-limited selection (left to right). The red dashed line in all panels shows the SDSS DR7 target magnitude limit $r'_{\rm lim}$ at $z = 0.1$. The red shaded region shows the extent of $r'_{\rm lim}$ for $0.09 \leqslant z \leqslant 0.11$. *Top panels:* the number density. Scale is not consistent between panels. *Bottom panels:* the stellar mass distribution. For the simulations this is the intrinsic stellar mass within the aperture. For SDSS this is the VESPA stellar mass estimates.

thereof, does not imply that either technique is more robust, but simply allows us to highlight the differences between our approach and an SED fitting approach.

5.5.1 SDSS Selection

We first selected all MGS galaxies in the redshift range $0.09 < z < 0.11$ where the redshift confidence was higher than 95%, which gave 76812 objects. We then removed those galaxies whose rest-frame wavelength coverage, with bad pixels removed, did not cover our fixed wavelength grid (see Section 5.2.4.4), yielding 66245 galaxies. Given our fixed wavelength grid we interpolated each spectrum (flux preserving; Carnall, 2017), de-redshifted and corrected for galactic extinction (Barbary, 2016b) using the Schlegel et al. (1998) galactic dust maps for each SDSS plate combined with the O'Donnell (1994) extinction curves ($R_V = 3.2$, where $R_V = A_V/E(B-V)$).

5.5.1.1 Aperture Correction

SDSS spectra are taken through a 3 arcsecond diameter fibre, which corresponds to 6 pkpc at $z = 0.1$. In order to apply our model, trained on galaxy spectra generated using a 30

pkpc aperture intended to mimic a petrosian aperture, we chose to scale up the observed fluxes by the mean of the difference between the fiber and petrosian magnitudes in the *observer* frame g and r bands (henceforth g' and r'),

$$S = 10^{0.2 \times ([M_{g'}^{\rm fiber} - M_{g'}^{\rm petro}] + [M_{r'}^{\rm fiber} - M_{r'}^{\rm petro}])} \ ,$$

where S is the flux scaling factor. After these corrections, the magnitude distribution on the $g - r$ plane of the selection at this stage can be seen in the top panel, third from left, of Figure 5.15. An alternative to scaling up the observational fluxes would have been to generate spectra from the simulations using a mock fibre aperture. Unfortunately, as discussed in Section 5.2.3.1, on these small scales resolution effects become important.

5.5.1.2 Colour Selection

We then used rest frame g and r magnitudes to perform a 2D selection on g and r band magnitude simultaneously (without replacement), in order to match the same 2D distribution from the combined Illustris and EAGLE samples (see the first two panels of Figure 5.15). SDSS spectra have a target apparent magnitude limit[26] of $r' < 17.77$, which corresponds to an absolute magnitude of -20.61 at $z = 0.1$; a large proportion of our simulated galaxies lie below this threshold, so we are limited to matching the distribution above this constraint (as shown by the red dotted line in all panels of Figure 5.15). Selecting galaxies above this threshold with matched magnitudes gives us a sample of 10 000 galaxies. It is clear from Figure 5.15 that the g against r distribution for SDSS galaxies deviates from 1:1 more so than the simulations, motivating the 2D selection. The selection based on the simulation broadband magnitudes is to ensure that, when used as features for the model, the spectra remain 'in-bounds' to some extent, i.e. are not outside the range of input training data. It is true that we *a priori* select observed galaxies with good spectral agreement with our simulations in a broad-band sense, however the details of the higher resolution spectra can still differ substantially. We have tested that our models do not fail dramatically on out-of-bounds SDSS data, however a more thorough test with simulated out-of-bounds performance is left for future work.

[26] https://classic.sdss.org/dr7/

In Appendix 5.8.3 we show how t-SNE can be used to evaluate the synthetic gap between the synthetic and observed spectra.

5.5.2 VESPA Star Formation Histories

The Vespa SFH catalogue predicts star formation histories with varying resolution depending on the quality and completeness of the input data, with a maximum resolution of 16 bins, though a resampled SFH at this higher resolution is also provided. We use this resampled SFH throughout the comparison, though caution that this does not necessarily represent the best fitting history. Vespa also provides predictions using the SPS models of both Bruzual & Charlot (2003) and Maraston (2005). The choice of model leads to significant differences in the predicted SFH, which highlights the effect of modelling choices on the inferred SFH. We use the Vespa results that use BC03 models assuming a Chabrier IMF, whilst noting that these will not necessarily lead to consistent predictions compared to the more recent FSPS models used in our model training, and do not include nebular emission. Using the more recent FSPS model is justified since the improved spectral modelling will lead to galaxies with more comparable intrinsic properties, such as the SFH, particularly since our selection is magnitude-matched to the SDSS sample. We leave a comparison of the effect of SPS model choice to future work.

SDSS DR7 spectra are measured within fiber apertures of 3" diameter. Vespa SFHs are corrected for this by scaling the entire normalisation (*i.e.* the mass in each bin) by the offset between the fiber and petrosian z-band magnitudes (Tojeiro et al., 2009),

$$M_{*,fiber} = \frac{M_{*,total}}{10^{0.4(z_f - z_p)}} \quad .$$

where z_f and z_p are the fibre and petrosian z-band magnitudes, respectively[27].

5.5.3 SDSS Predictions

Figure 5.16 shows SFH predictions fromVespa and our Illustris and EAGLE models (trained on dust-obscured spectra, with noise resampled $\times$ 3) for four example SDSS galaxies. We emphasise that neither our model nor the Vespa predicted histories represent

[27]In Tojeiro et al. (2009) the equation for the stellar mass correction contains an error; it is reproduced here correctly

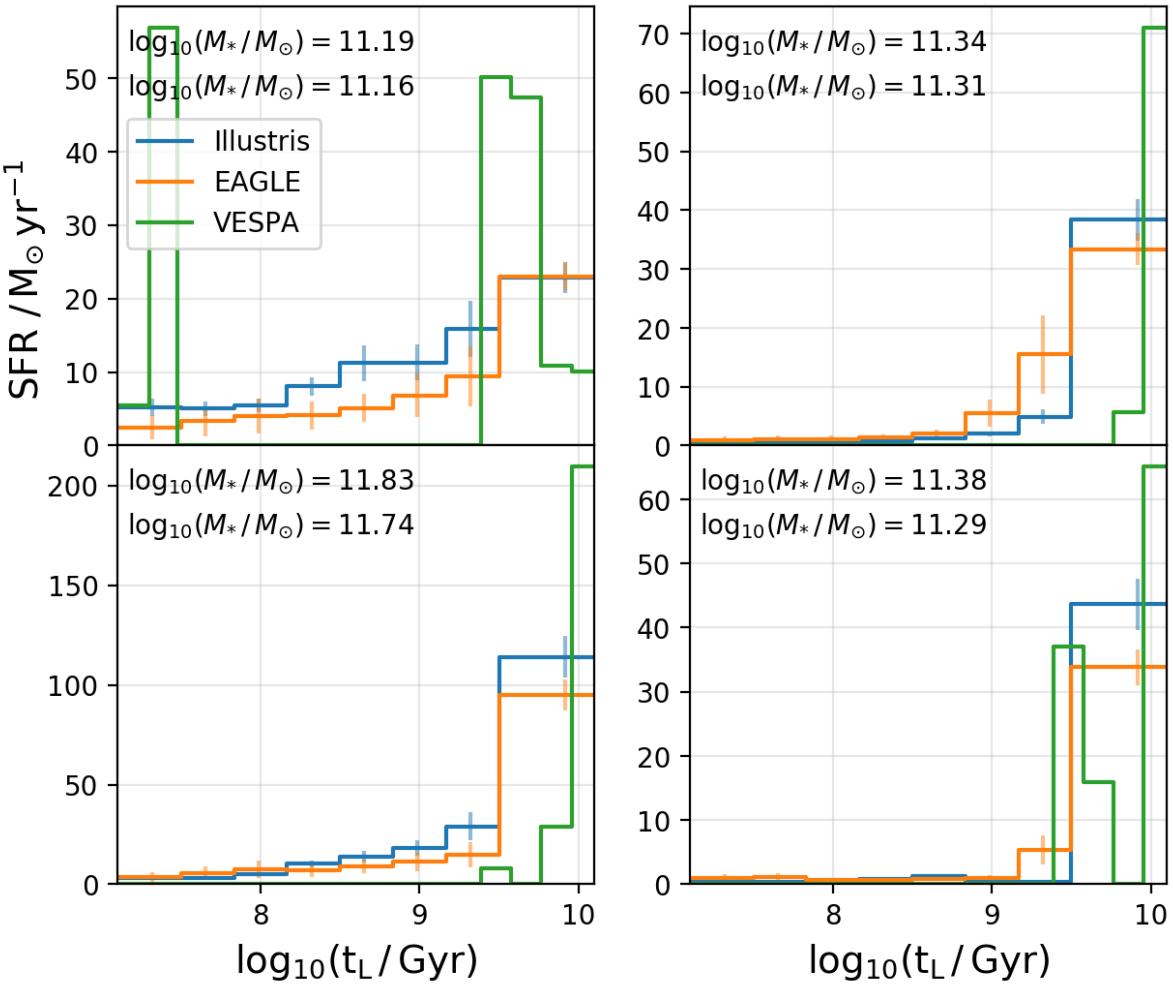

Figure 5.16: Four example SFHs from VESPA, alongside predictions for the same SDSS galaxies from the EAGLE and Illustris models (trained on dust-obscured spectra with noise, resampled × 3). We show histories with total predicted masses from the Illustris model closest to the estimated VESPA total masses. Uncertainties are estimated from the observational and modelling errors, described in Section 5.4. Our models trained with EAGLE and Illustris predict similar shaped histories, with smoother evolution than VESPA.

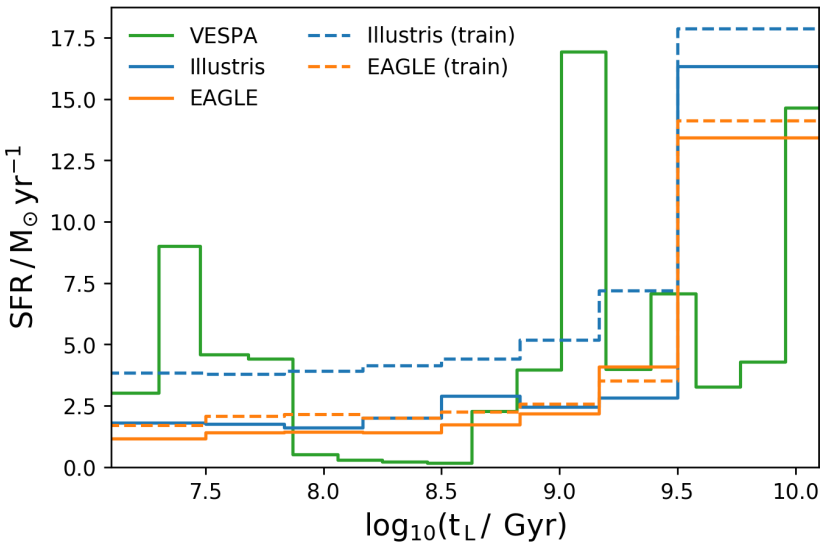

Figure 5.17: Mean predicted SFH for the SDSS selection from VESPA, and our Illustris and EAGLE models (including dust and noise, resampled × 3).

the 'true' SFH, but are shown simply to highlight the differences. Our model SFHs are much smoother than those predicted from VESPA, which predicts more stochastic, bursty histories.

Our observational selection is neither mass nor volume complete, so it is not possible to make a fair evaluation of the population SFH or cosmic star formation rate density as a function of time. However, we can plot the median SFH from each model for this selected sample to better understand the ensemble prediction, shown in Figure 5.17. VESPA predicts two large peaks in the SFR distribution at $\sim 200\,\mathrm{Myr}$ and $\sim 1\,\mathrm{Gyr}$, whereas our model predictions for Illustris and EAGLE have smoother, decreasing behaviour, peaked in the earliest bin.

Figure 5.17 also shows the median input SFH from the simulation for a sample magnitude-matched to the observations, which can be thought of as the effective 'prior' on the SFH distribution. The predicted distributions are similar, though not identical, to the training distributions, which suggests that the prior is highly informative, as expected, but does not dominate.

We also estimate the final stellar mass of each galaxy from the SFH by assuming an age dependent recycling fraction (estimated using python-FSPS; Foreman-Mackey et al., 2014). Figure 5.18 shows our estimates obtained from the EAGLE and Illustris models compared to the VESPA estimates. Both models return similar stellar masses to VESPA,

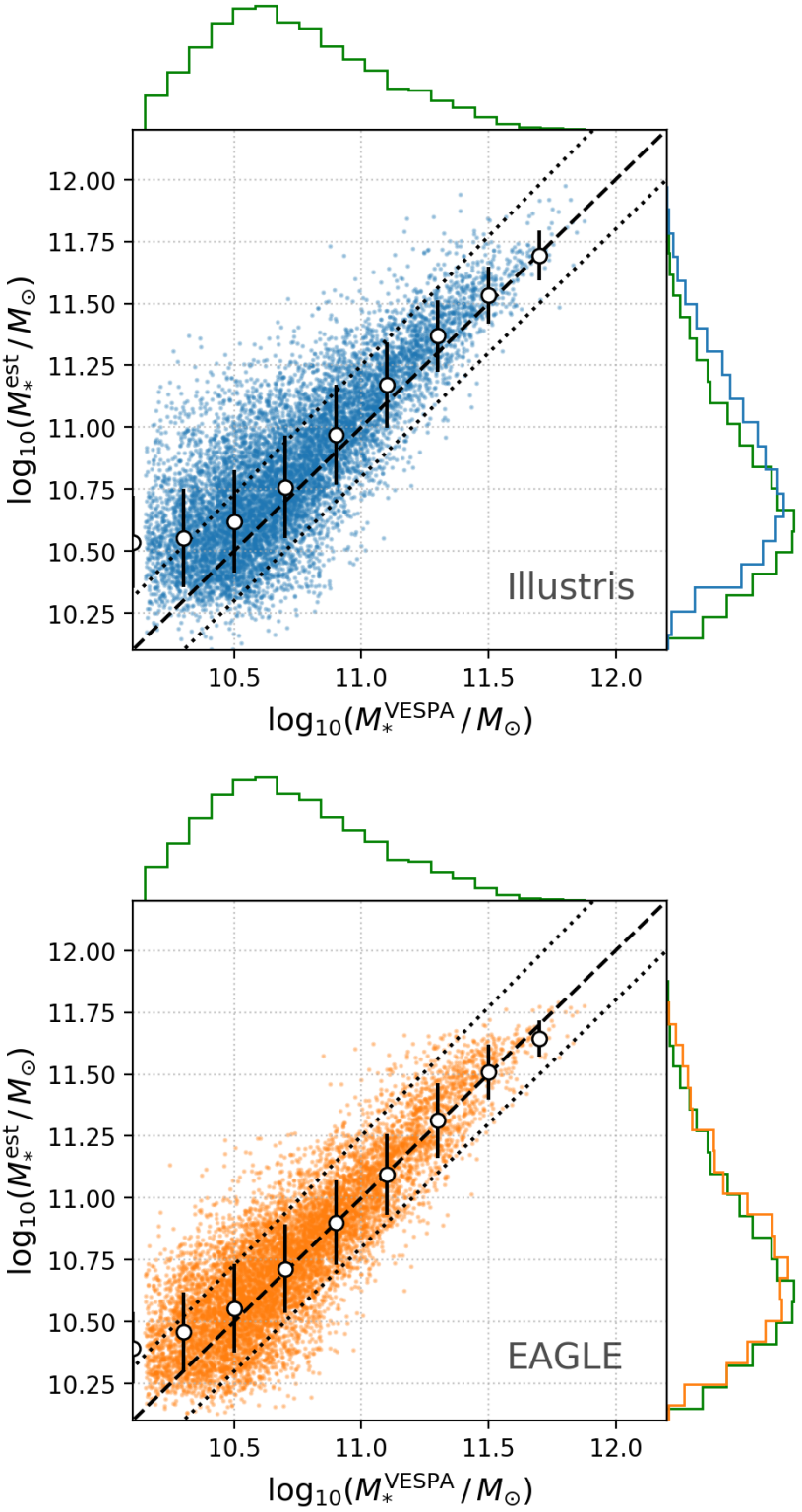

Figure 5.18: Estimated final stellar masses from the predicted SFH in the Illustris (top, blue) and EAGLE (bottom, orange) models, assuming an age dependent recycling fraction, compared to those published in the VESPA catalogue. The black dashed line shows the one-to-one relation, and the dotted black lines show ± 0.25 dex offset. The white points show the binned median and 1σ scatter. The histograms at right show the marginal distributions of estimated stellar masses; the histogram for the VESPA distribution (green) is shown at top, and at right for comparison. The mass estimates are very similar to those obtained from VESPA down to $\log_{10}(M_*\,/\,M_\odot) \sim 10.5$, with little scatter.

within ~ 0.25 dex for the majority of galaxies, and there are no mass dependent trends down to $\log_{10}(M_* \,/\, M_\odot) \sim 10.5$; there is a floor to the predicted masses, due to the lack of simulated galaxies with such low masses in the magnitude-selected sample.

Finally, Figure 5.19 shows the median predicted SFH from each model, binned by total predicted mass (from the Vespa model). For EAGLE, all four bins show very similar behaviour, in both the median and the 16^{th}-84^{th} percentile spread around it, except in the earliest bin, which has higher SFR for more massive galaxies. Illustris, in contrast, predicts a peak in the SFH at intermediate ages for lower mass galaxies, giving younger average stellar ages.

5.6 Discussion

We have demonstrated a new approach to estimating star formation histories using cosmological simulations, combined with detailed synthetic spectral modelling, to train a convolutional neural network. This approach is subject to different systematics and modelling assumptions compared to traditional SED fitting, which we discuss in greater detail here, as well as possible extensions in future work.

5.6.1 Cosmological Simulations

One of the obvious limitations to using cosmological simulations as a training set is that our understanding of galaxy formation is incomplete, and as such cosmological simulations are not truly representative of actual galaxies, neither individually or in ensemble, which can impact the predicted SFHs. More realistic modelling is an obvious remedy, though this is already a fundamental aim of galaxy evolution studies.

One way of evaluating the predicted population SFH distribution is to look at the evolution of the cosmic star formation rate density (CSFRD), which in hydrodynamic simulations has been shown to be consistently in tension with observational constraints at cosmic noon ($z \sim 2$) (Somerville & Davé, 2015). Key distribution functions in EAGLE and Illustris of point-in-time properties, such as stellar mass and star formation rate, are also in tension with both observations and each other at high redshift. Semi-analytic models are able to match these distribution functions better at a range of redshifts (*e.g.* Henriques et al., 2015; Clay et al., 2015), but do not resolve the stellar populations.

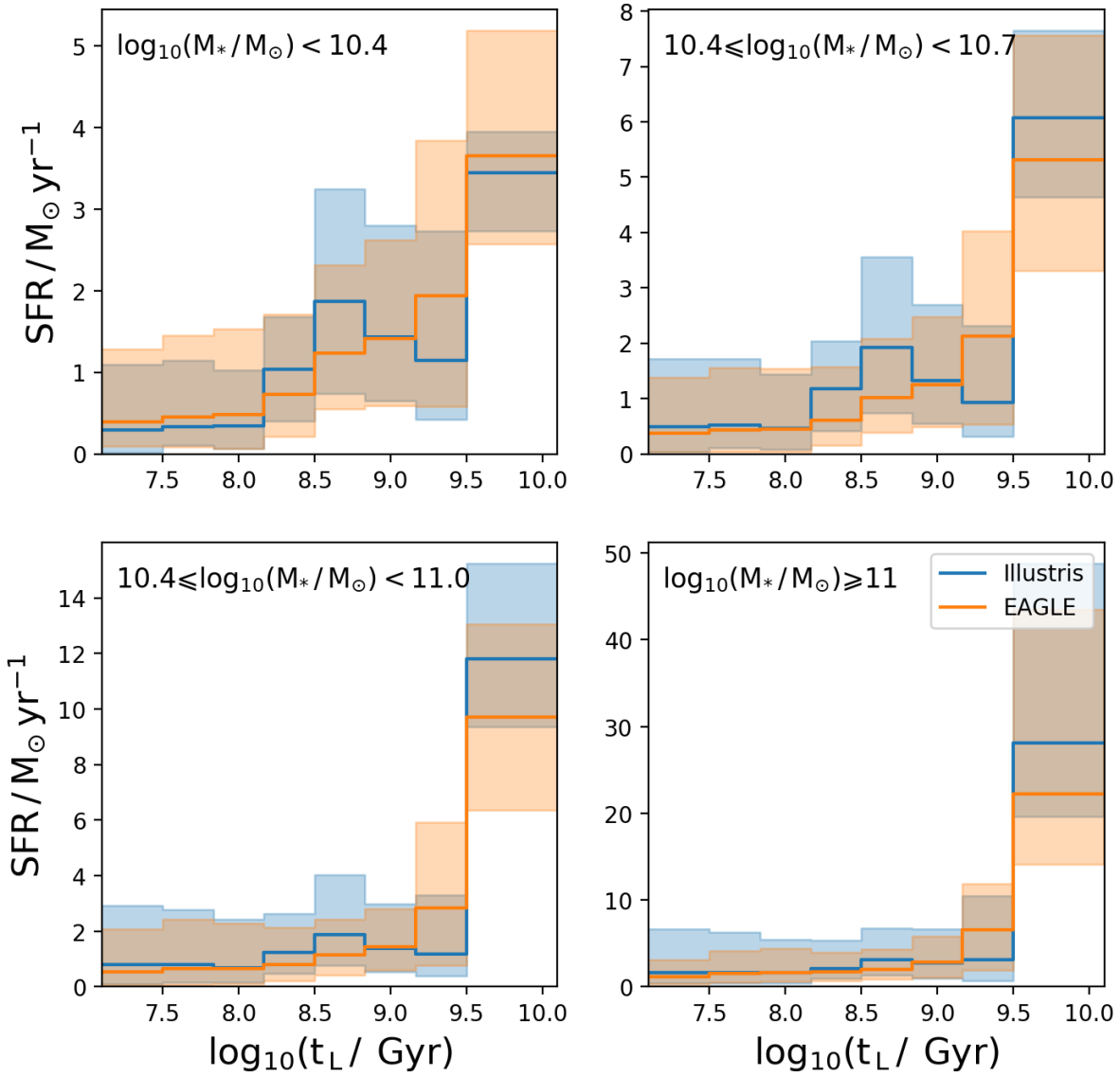

Figure 5.19: SDSS predictions from EAGLE and Illustris split by VESPA predicted total mass. The lines show the median, and the shaded region the 16^{th}-84^{th} percentiles. EAGLE and Illustris SFH predictions for low mass galaxies are significantly different, with Illustris predicting a younger average population.

Incorrectly predicted galaxy properties also impact spectral modelling where it is physically motivated. One physical property that has a large impact on our dust model is the central cold gas mass; both EAGLE and Illustris have been shown to underestimate this mass, to differing extents (Crain et al., 2015; Genel et al., 2014). We find that the average star forming gas mass in Illustris is higher than in EAGLE for our selected galaxies; in EAGLE there are a significant number of galaxies with zero star-forming gas, which gives zero attenuation in our dust model ($\gamma = 0$). This leads to higher average attenuation for Illustris galaxies, however this is cancelled out to some degree by the higher median SFR in Illustris over our mass range (see Figure 5.2), which leads to higher intrinsic luminosities. This could possibly explain the good agreement in optical colours with observations presented in Trayford et al. (2015) and Genel et al. (2014), despite the differing star-forming sequence behaviour between the simulations.

Trayford et al. (2015) find that, using a very similar dust model to that used in this work, EAGLE galaxies over the stellar mass range $10^{10.5} < M_* / M_\odot < 10^{10.8}$ exhibit a stronger bimodal colour distribution than that seen in observations from the GAMA survey (Taylor et al., 2015), with higher fractions of blue galaxies. This strong bimodal behaviour remains when the authors use an orientation dependent dust model. The colour distribution may be related to the lower passive fractions ($\sim 20\%$) seen in this mass range compared to observations (Schaye et al., 2014). Such trends will affect predictions from the EAGLE-trained model, since its 'prior' for the SFR distribution will be skewed towards more star forming objects that may not be representative of the true SFR distribution of galaxies. Similar arguments can be made for the Illustris predictions, where the SFR distribution has a higher normalisation for intermediate masses. We do not find that the model stellar mass estimates for SDSS galaxies show significant biases compared to VESPA, but the mass is dominated by the wide early bin. We could of course have selected SDSS galaxies with similar stellar masses, but these may have been out-of-bounds in spectral space. We conclude that improved physical and spectral modelling in the simulations to match the magnitude - stellar mass relations would improve our predictions.

5.6.2 Spectral modelling

The difference between synthetic spectra and observed spectra, known as the *synthetic gap*, can lead to significant biases in predicted histories. More sophisticated approaches

to modelling the dust could reduce this gap. Dust models that take in to account the geometry of the gas and stars within the system show better agreement with observed colour distributions (Trayford et al., 2015; Davé et al., 2017). The most sophisticated approach employs 3D Monte-Carlo radiative transfer (RT), which treats absorption and anisotropic scattering by dust, as well as thermal re-emission and dust heating, in a self-consistent way. This approach has been applied to the EAGLE simulations using the SKIRT code, to calculate the FIR and dust properties of the galaxy population (Camps et al., 2016; Trayford et al., 2017); they find a better match to observed local colour distributions compared to screen models. Introducing such line of sight dependence on the attenuation is expected to reduce the correspondence between the simulated spectra and the underlying SFH, equivalent to reducing the information content of the spectra for learning our target property, the SFH. This may lead to greater uncertainties in the derived SFHs; we will explore the effect of this in future work.

In Appendix 5.8.3 we briefly explore the use of t-distributed Stochastic Neighbour Embedding (t-SNE) to evaluate the similarity of our synthetic spectra to the observations. Whilst the results are good for visualisation purposes, this method is particularly sensitive to the choice of hyperparameters, such as learning rate and complexity. Masters et al. (2015) demonstrate how self-organised maps can be used as an alternative means of addressing similarity in multi-dimensional feature spaces, whilst requiring fewer free parameters. We plan to use this in future work as an alternative, potentially more robust way of assessing the synthetic gap.

5.6.3 Machine learning approach

We use a simple cut in stellar mass to select our training sample, which we found does not lead to overfitting of low mass galaxies despite the steepness of the GSMF (see Section 5.3.1.5). It is unclear whether the lack of overfitting to low mass objects would extend to lower stellar masses, however the results presented here are promising. Predictions for rare objects could also be improved by using larger volume simulations and/or 'zoom' resimulations of biased regions, to increase the sampling of extreme objects, though this would negate the advantage gained from using a representative sample.

We rely on cosmological simulations for training data due to the small number of galaxies

(∼20) for which resolved, reasonably confident measurements of the true SFH are known. Such objects are also mostly in the local universe, restricting any predictions to this period. However, with ever increasing samples locally, including from from integral field unit (IFU) spectrographs (Bundy et al., 2015; González Delgado et al., 2017), it may soon be possible to train a machine on high resolution observational data in order to predict the SFH of galaxies with only unresolved data on a larger number of objects.

Our approach relies on a fixed grid of input features. Where observational data do not cover this wavelength range we currently ignore them. An alternative to this would be to impute missing features, for example through interpolation.

5.6.4 Future Extensions

A unique aspect to our approach is that it can take advantage of the detailed modelling of complex, non-linear processes in the simulations to infer more physically motivated SFHs. This could also be extended to other quantities self-consistently predicted in the simulations, but not directly responsible for the optical emission. For example, halo mass could be used as a predictor, and the results compared to abundance matching approaches. We will explore this in future work.

A powerful complement to using spectroscopic features would be to use multi-wavelength photometry, such as that available in the CANDELS fields. However, convolution across this smaller, heterogeneous feature set would be inappropriate; using a tree based or fully connected network would lead to better performance, both computationally and predictively. We could then compare our results to those obtained via SED fitting on photometry, using different codes such as SpeedyMC (Acquaviva et al., 2015) and other alternatives such as Prospector (Leja et al., 2017) and BEAGLE (Chevallard & Charlot, 2016). This will clarify how the biases and projected uncertainties of the two techniques compare, and help us make a final recommendation on improved star formation histories from multiple methods. Tree based methods also provide information on feature importance, by equating importance with depth of features in the tree. Deep learning based approaches necessarily obscure the relationship between the predictors and the features through the complexity of the built network, which makes it difficult to extract feature importance.

5.7 Conclusions

We have used convolutional neural networks (CNN) to learn the relationship between galaxies spectra and their star formation histories (SFH), using synthetic spectra generated from two cosmological hydrodynamic simulations, EAGLE and Illustris, as our training data. Our findings are as follows:

- The CNN is capable of recovering the SFH of test galaxies to high accuracy (SMAPE = 10.9%), despite the presence of dust and noise, and with no significant bias with stellar mass, SFR or stellar-metallicity.
- We estimate the uncertainty in our predictions from observational errors and modelling errors, and use these in combination to provide a realistic error budget on unseen data. Modelling errors dominate for both dust-obscured and intrinsic spectra.
- We demonstrate the good generalisation properties of the technique by applying a model trained on one simulation to simulated data from another, obtaining good accuracy (SMAPE = 14.4% for the dust-attenuated Illustris model applied to EAGLE data) even on these unseen spectra. The model also recovers the star-forming sequence of the input data, which suggests it is not overfitting to a particular simulation.
- We apply our models to a magnitude matched sample of SDSS DR7 spectra and compare to the SFHs from the VESPA catalogue. The model predicts smoother SFHs, influenced by the 'prior' distributions from the simulations, whilst recovering consistent total stellar mass predictions.
- When applied to our SDSS selection, the Illustris-trained model predicts younger average stellar ages for low mass galaxies ($\sim 10^{10} M_* \,/\, M_\odot$) than the EAGLE-trained model. For higher mass galaxies ($\sim 10^{10} M_* \,/\, M_\odot$) both models predict similar SFH (and hence age) distributions.

Table 5.1: Fitted parameters for the observational and modelling errors. The first two columns state the bin edges in log-lookback time. m_2, m_1 and c give the second order polynomial fit parameters to the observational error. $\sigma_{\rm model}$ gives the 1σ spread in a normal fit to the fractional residual distribution.

Bins $[\log_{10}(t_L)]$		m_2	m_1	c	$\sigma_{\rm model}$
0.00	7.50	0.31	0.17	0.00	0.22
7.50	7.83	0.36	0.10	0.01	0.19
7.83	8.17	0.39	-0.02	0.02	0.17
8.17	8.50	0.32	0.03	0.02	0.16
8.50	8.83	0.39	0.01	0.03	0.21
8.83	9.17	0.58	-0.15	0.06	0.21
9.17	9.50	0.63	-0.12	0.09	0.24
9.50	10.10	0.63	-0.47	0.16	0.09

5.8 Appendix

5.8.1 Correlation matrices

The correlation matrix can be inferred from the covariance matrix of the spectral errors, and shows the interdependence of each bin as a result of changes to the input spectra. It is given by

$$r_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j} \ ,$$

where $r_{ij} \in [-1, 1]$. Figure 5.20 shows the correlation matrix for each galaxy shown in Figure 5.8; the corresponding indexes are quoted in the top right of each panel. In general, adjacent bins show the highest correlation, as expected since they are constrained by similar spectral features. More distant bins tend to show anti-correlation, which may be due to the stellar mass constraint; where SFR increases in one bin, it is reduced in others so that the total stellar mass is reproduced.

5.8.2 Error Tables

In Section 5.4 we describe our method for estimating the uncertainty in the SFH predictions from *observational* and *modelling* errors. In Section 5.4.1 we fit second-order polynomials to the mean observational error distribution in each bin, for dust-obscured spectra,

$$e_{\rm exp} = m_2 x^2 + m_1 x + c \ .$$

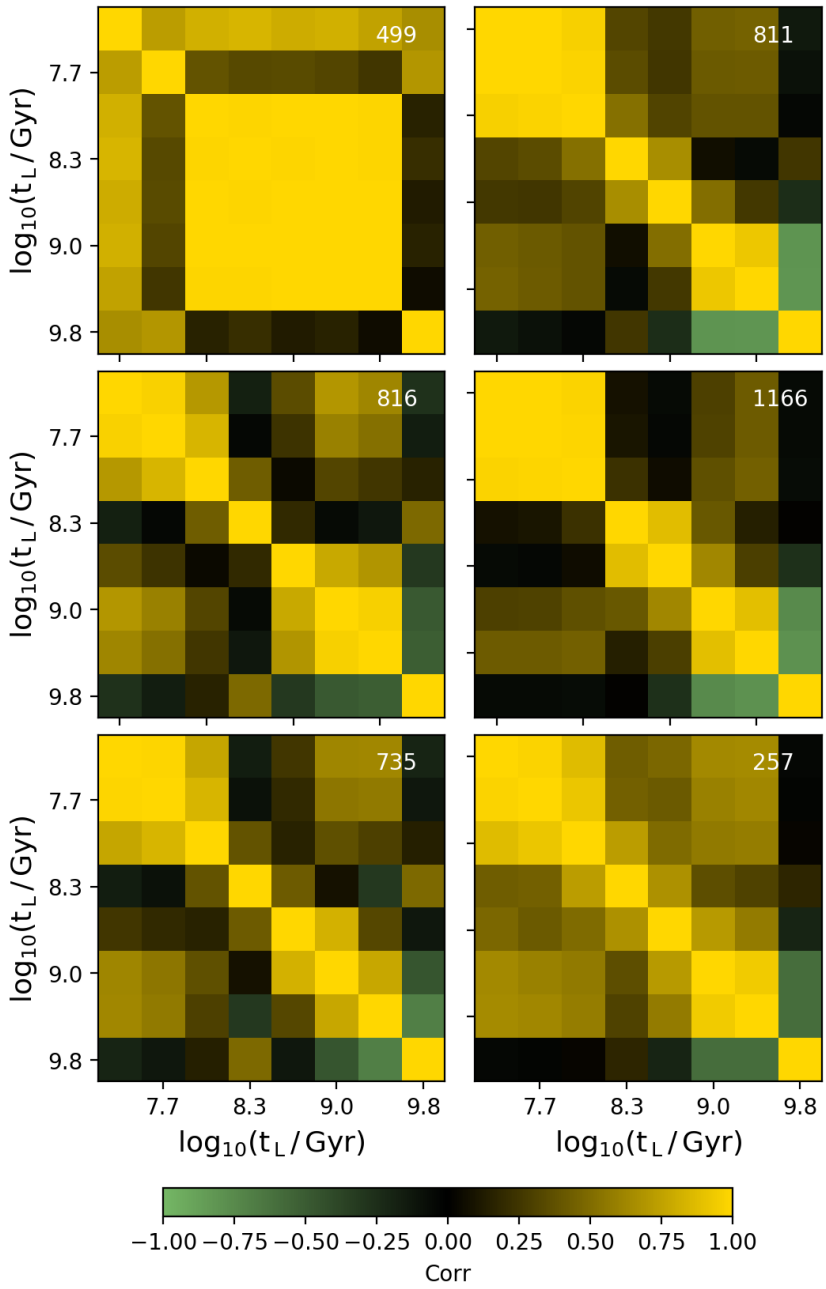

Figure 5.20: Correlation matrix from spectral errors, for the six galaxies shown in Figure 5.8 (the corresponding indices are printed in the top right corner of each panel). The colour scale varies through yellow, black and green, which show positive, neutral and negative correlation, respectively.

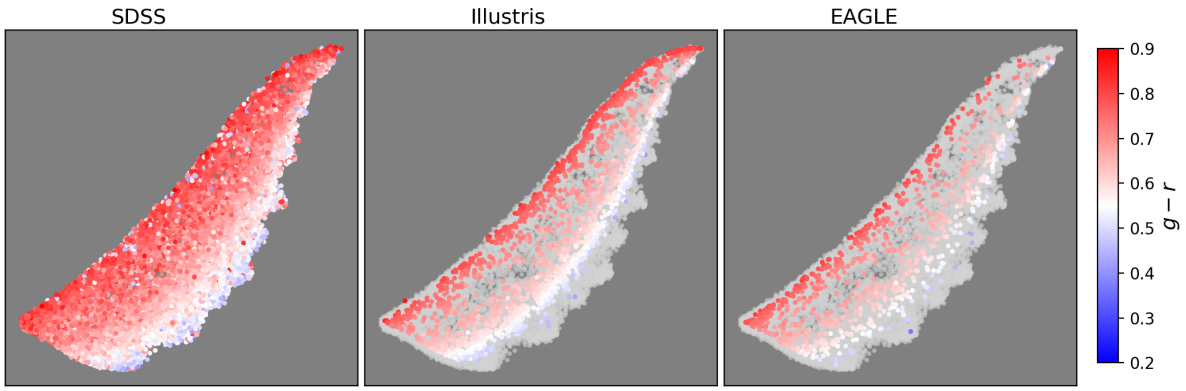

Figure 5.21: t-SNE plot applied to spectra from the SDSS selection (left panels) and the Illustris (middle panels) and EAGLE (right panels) selections. Each point represents a single galaxy spectrum. Nearby points in this 2D space have high spectral similarity. Each distribution is coloured by $g - r$ colour. The SDSS selection is shown in the background in light grey for the middle and right panels for comparison.

The fit parameters are shown in Table 5.1. In Section 5.4.2 we fit the fractional residual distribution with a normal; the 1σ spread is quoted in Table 5.1. To obtain the 1σ modelling error simply multiply the predicted SFR in each bin by σ. To estimate the *total* error in each bin, add the observational and modelling errors in quadrature. The distribution of fractional residuals is slightly non-symmetric, resulting in an over-estimate of the average error; we have tested the effect of this by measuring the fraction of the true SFH that lies within the errors, and found that this is the case for approximately 70% of cases, close to the 1σ definition.

5.8.3 t-distributed Stochastic Neighbour Embedding

In order to generate robust predictions using a supervised machine learning model, one needs confidence that data used to train the model are representative of the data to which it is to be applied.[28] Synthetic spectra will always exhibit a bias compared to observational spectra, known as the *synthetic gap*; where it is large it can limit the applicability of learning algorithms trained on synthetic data to observations. To evaluate the synthetic gap we use t-distributed stochastic neighbour embedding (t-SNE) (Maaten & Hinton, 2008), a technique for reducing high dimensional data down to a lower number of dimensions whilst preserving the multi-dimensional distance, for visualisation purposes (Wattenberg et al., 2016).

[28] One approach, proposed by Cohn & Battaglia (2019) in the context of galaxy cluster mass estimation, is to compare inferred correlations between observables in the simulations to those in actual observables.

Figure 5.21 shows the result of running t-SNE on the observationally matched sample of synthetic spectra, and the observations themselves. The EAGLE and Illustris spectra are clustered in very similar regions of the two dimensional space, which suggests they exhibit very similar spectra. We emphasise that t-SNE evaluates the synthetic gap across the whole of the feature space; close correspondence in this space suggests very close spectral similarity. The observational results overlap with the simulations well, though there are certain regions, particularly at the edges of the 2D distribution, where they cluster separately from the simulation distributions, suggestive of a synthetic gap. Figure 5.21 shows each distribution coloured by $g - r$ colour; where the simulations and the observational spectra do not overlap in this distribution tends to be in the extremes of the colour distribution. This may be due to the limited volume of the simulations used for training ($\sim 10^6\,\mathrm{Mpc}^3$), which will sample fewer extreme objects, such as those in dense cluster environments. More sophisticated approaches to spectra generation (e.g. full radiative transfer) will enhance the physical realism of the synthetic spectra, and may also reduce this synthetic gap (see Section 5.6).

6 Conclusions

In this thesis I have used numerical and machine learning approaches to study star formation in galaxies throughout cosmic time, and its environmental dependence. In chapters 3 and 4 I focused on galaxy protoclusters, the high redshift progenitors of galaxy clusters. Protoclusters are rare and extended, which makes it difficult for both simulators and observers: simulators must use large volumes to capture enough protoclusters for a statistically significant sample, and observers require similarly large sky coverage, as well as good redshift estimates to constrain the galaxies to the protocluster. I used both semi-analytic and hydrodynamic simulations; combining these different numerical techniques is necessary, to understand protoclusters in detail individually, and to study a large enough population to do statistics, respectively. In the final chapter I present a novel method for inferring the star formation history (SFH) of a galaxy, using machine learning methods coupled with state-of-the-art numerical simulations. This method uses an alternative way of looking at star formation at high redshift, utilising the abundance of good quality data available for nearby galaxies to infer the star forming-activity of their progenitors.

Chapter 3 focuses on the identification of protoclusters, and their characterisation in terms of descendant halo mass, using the L-Galaxies Semi-Analytic Model. I present the first measurement of protocluster shapes as revealed by different galaxy tracers. I also use the concepts of completeness and purity throughout this chapter to better understand the relation between observed and true protoclusters. First, I derive an optimum search aperture for protoclusters, and find that $R \sim 10$ cMpc maximises both completeness and purity of the galaxy population, irrespective of selection and redshift, for $2 \leqslant z \leqslant 10$. This aperture also best identifies overdensities surrounding AGN, since typically these are not located centrally within the protocluster. Finally, I combine these results into a single, comprehensive criterion for identifying protoclusters from galaxy overdensities. The procedure returns the probability that a given measured overdensity is one of four different classifications: field, protocluster, part-protocluster, or protocluster-field. This approach does not assume that a given measured overdensity is centred on the protocluster, nor that it has captured all of the protocluster galaxies, and so gives more realistic protocluster probabilities and descendant masses. Applying the method to

historical candidate protoclusters from the literature, I conclude that many are not as highly significant as first thought.

In Chapter 4 I use the full hydrodynamic simulation C-Eagle, a series of zoom simulations of cluster environments with a range of descendant halo masses. These simulations follow the evolution of baryons and dark matter in galaxies self-consistently, and allow us to study the history of star formation in galaxy clusters in detail. I present a study of the star-forming sequence (SFS) in protoclusters, which describes the average star formation rate at a given stellar mass. At low redshift the SFS shows a clear dependence on environment, but it is unclear what this environmental dependence is at high redshift, and whether it has a positive or negative effect on galaxy star formation. I find that the SFS has a very similar form in protocluster environments to the field, but with some notable differences. At $z > 3$ the specific-star formation rate (sSFR) is significantly discrepant from the field, as evidenced by a Kolmogorov-Smirnov test, with a higher normalisation that is particularly evident at $z \sim 6$. The sSFR distribution is similarly discrepant at $z \sim 1.5$, where the collapse of the most massive clusters is well underway.

Passive galaxies are ubiquitous in nearby galaxy clusters, but their presence at high redshift, and any environmental dependence of the quenching mechanism, is still uncertain. In Chapter 4 I study the passive fraction in C-Eagle protoclusters. I find that the fraction of galaxies in protoclusters is similar to the field in the model, in tension with observational constraints at $z \sim 2$ which show higher passive fractions in protoclusters. In the simulations I find that the dense group environments have lower passive fractions than the intergroup, a surprising result that suggests groups have a positive impact on SFR at $z \sim 2$. Even more surprising is that the intergroup galaxies have higher passive fractions than the field, which suggests these regions, whilst not within the highest overdensities in the protocluster, are still (negatively) affected by the large scale overdense environment.

It has typically been assumed that protoclusters would preferentially host galaxies with active galactic nuclei (AGN). However, the presence of this relationship, and its physical cause, is still uncertain. In Chapter 3, using the L-Galaxies two-component AGN model, I conclude that AGN are complicated tracers of protoclusters; using the concepts of completeness and purity again, at high redshift AGN have low completeness and high purity, but at lower redshifts the opposite is true; AGN are highly complete tracers, but

with low purity. I do not study the AGN-protocluster relation in C-Eagle explicitly in Chapter 4, however the SFS exhibits a clear turnover, which can be attributed to the onset of AGN feedback. This turnover mass show no difference between protoclusters and the field, which suggests the stellar mass at which AGN feedback kicks in, and any associated physical cause *e.g.* higher gas mass for accretion, is not environmentally dependent.

The turnover in the SFS in C-Eagle evolves to lower stellar mass with increasing redshift, which is in tension with recent observational results which show the opposite redshift dependence, at least up to $z \sim 3$. However, I note that the form of the SFS measured is sensitive to the lower mass limit - many observational studies that fit a single power law quote a shallow slope, which I suggest is due to the fact that they are only probing the high-mass end of a two-part relation. I also measure the scatter around the SFS, and find that satellites in protoclusters show increased scatter compared to the field. The scatter in the centrals relation, however, shows no significant environmental dependence, which suggests the effect of environment is limited to interactions on small, inter-halo scales, rather than intra-group scales.

Finally, in Chapter 5 I present a novel way of predicting galaxy SFHs. I use the outputs of two cosmological simulations, Eagle and Illustris, with detailed modelling of their SEDs, to learn the relationship between a galaxies SFH and its spectra. This is essentially a supervised regression problem, and I demonstrate good performance on test sets from each simulation. I also show the good generalisation properties of the method by testing models trained on a given simulation to spectra and SFHs from another simulation, and provide estimates of the experimental and modelling errors.

The cosmological simulations can provide more realistic and informative priors for the SFH that self-consistently take into account the cosmological evolution of all components that contribute to the galaxies SED, such as the evolution of stellar and gas-phase metallicity, gas mass, and morphology. The cosmological simulations also provide information on the most common SFH for a given SED by default, something that in traditional SED fitting is not readily available. Our non-parametric form for the SFH also avoids many of the biases present when using simple parametric forms. Obviously, if the training simulation is not representative of the true universe you will achieve biased results, however this model dependent bias can be evaluated by using many different simulations, to predict on

a given observational dataset and to evaluate their generalisation properties.

I provide a practical demonstration of the method by applying it to spectra from SDSS DR7 (Abazajian et al., 2009), and compare to the VESPA catalogue (Tojeiro et al., 2007, 2009), finding much smoother histories, as well as consistent stellar mass estimates. This smoothness better fits measurements of the cosmic star formation rate density, whereas VESPA is biased to certain high-information simple stellar populations, leading to a more stochastic SFH. Interestingly, the Illustris trained model predicts younger stellar populations on average for lower mass galaxies, translated from the higher SFS normalisation in the original simulation.

6.1 Future Work

An obvious extension on the work in Chapter 3 would be to use an SED motivated selection function in the L-GALAXIES model. Modelling Lyman-α emitters and Lyman-break galaxies explicitly would allow us to make clearer comparisons with observational studies, such as SILVERRUSH and GOLDRUSH (Toshikawa et al., 2017; Higuchi et al., 2018). Another valuable extension would be to test the sensitivity of our results to the chosen cosmology, semi-analytic model and merger tree code. SHARK (Lagos et al., 2018) is an open source, modular SAM that runs on the merger trees from the phase space structure finder VELOCIraptor; running this on the Millennium simulation and comparing to our results would be a good test of the robustness and model dependence. Running new, large dark matter simulations would also be of interest, and is now computationally cheap thanks to new state-of-the-art numerical codes such as SWIFT (Borrow et al., 2018). Using a box size of order $(\sim 1\mathrm{Gpc})^3$ would produce a much larger sample of high-mass clusters for study and comparison with observations.

The C-EAGLE simulations represent a rich resource for future protocluster studies. Possible future avenues for research include investigating the AGN-protocluster connection in the EAGLE AGNdT9 model, and the spatial distribution of AGN in protoclusters. Parametrising the galaxy stellar mass function and star formation rate distribution function in protoclusters as a function of descendant mass would also be a valuable extension, allowing observers to directly compare their measured mass functions with the simulations. Together with collaborators I have developed a Bayesian fitting procedure for

these distribution functions that allows us to provide full posteriors on the fit parameters, that can be propagated through subsequent analyses. Additionally, I already mention in Chapter 4 that the formation and assembly of the brightest cluster galaxy in C-EAGLE is an interesting avenue for exploration due to the offset seen in Bahé et al. (2017), and a challenging new high-stellar mass frontier for testing the EAGLE physics model, particularly feedback in massive progenitor galaxies at high-z. Figure 6.1 shows early exploration work on this subject, displaying the formation and assembly times of the stellar mass in BCG progenitors. There is significant diversity between different descendant mass protoclusters.

The high-z universe presents a new frontier for both observers and theorists. Whilst the protocluster environment is of interest to study because of its descendant relation, the impact of environment at high redshift is not limited to these large-scale overdense environments. Indeed, I have shown in Chapter 4 that it is within and surrounding dense groups within protoclusters that environment plays the biggest role, and these groups can also exist in average density environments at high redshift. A more comprehensive study of the environmental impact is therefore warranted, and I have begun an investigation of this with collaborators. We use the same procedure as C-EAGLE to select zoom regions, but rather than selecting low redshift clusters we select a range of different overdensities at high redshift ($z \sim 4.5$). We can then combine these regions together with a similar procedure to the GIMIC simulations (Crain et al., 2009) in order to produce distribution functions with much larger dynamic range. Figure 6.2 shows the predicted UV luminosity function (generated using the detailed SED modelling developed in Chapter 5) using this procedure, along with the predicted depth and coverage of a number of upcoming space-based observatories. We are able to probe a much larger dynamic range than typical hydrodynamic periodic box simulations.

In Chapter 4 I suggest that the Hα-SFR correction in Kennicutt Jr & Evans II (2012) may be biased by ignoring the contribution from binary stars, as well as the effect of low metallicities at high redshift. With collaborators I am currently working on a new correction using the BPASS models, incorporating the evolution of the average galaxy metallicity from numerical models such as EAGLE. We plan to apply this to the MAHALO protoclusters (Shimakawa et al., 2017a, 2018) and derive new SFS relations that may be in reduced tension with the SFS from simulations.

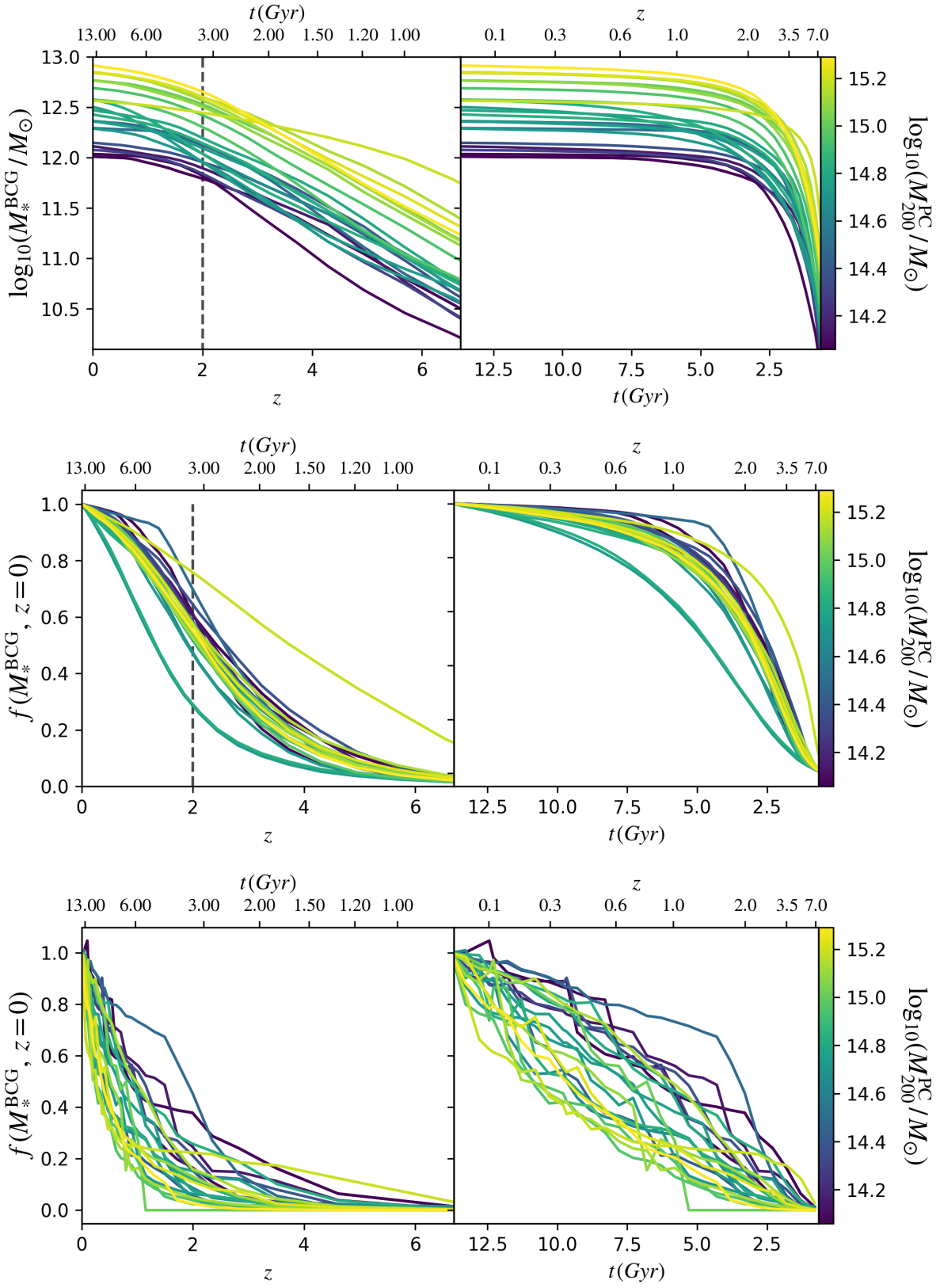

Figure 6.1: Evolution of the stellar mass of the BCG progenitors in the C-EAGLE sample. *Top:* total stellar mass formation time. *Middle:* fractional formation time. *Bottom:* stellar mass assembly time in the main branch progenitor.

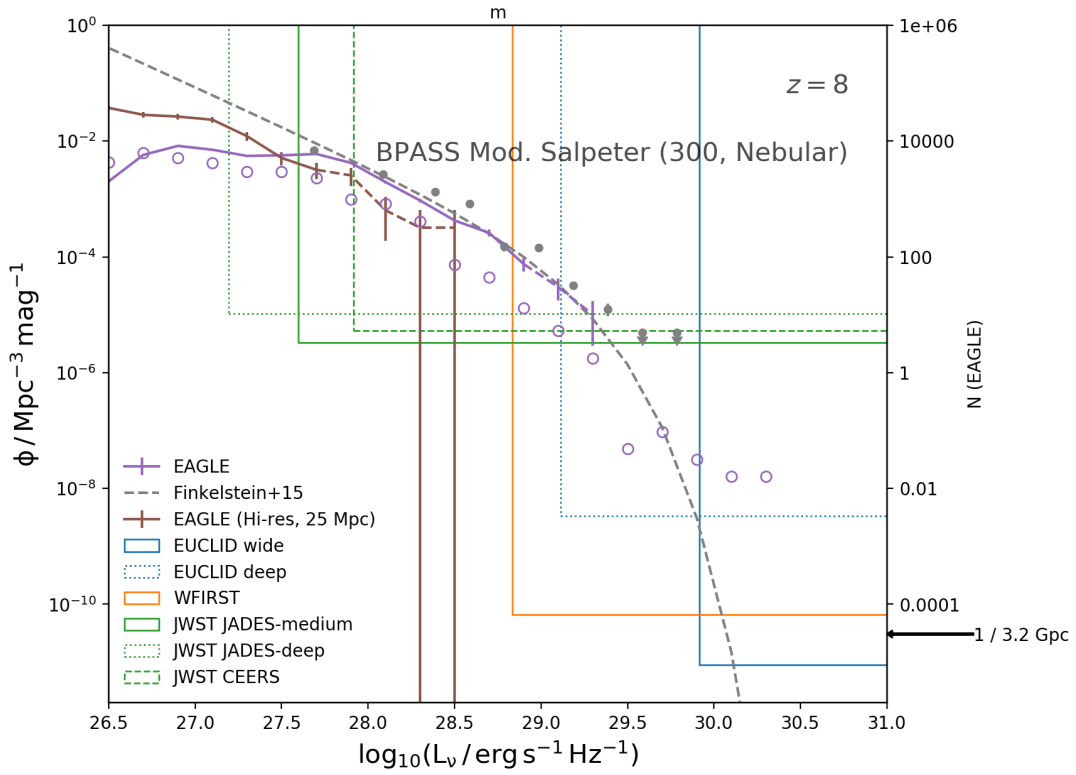

Figure 6.2: The UV lumimosity function at $z = 8$, with current observational constraints from Finkelstein et al. (2015); Bouwens et al. (2015), and forecasts for coverage from upcoming observatories. Current constraints from the fiducial and high-resolution EAGLE simulations are shown in purple and brown, respectively. Predictions from the combined high-redshift sample are shown with empty purple points; the resimulations extend the dynamic range of the UVLF considerably over the periodic volumes.

The machine learning method presented in Chapter 5 opens up a number of opportunities for future extensions. The simplest involve using the latest simulations, such as the Illustris-TNG model and the new SIMBA simulations. Other avenues include using more sophisticated SED modelling approaches, such as full radiative transfer, extending to higher redshift, and using photometric data rather than sull SEDs. Since photometric data does not represent a continuous feature set, convolution across this heterogeneous feature set would be inappropriate; tree based approaches would be a more suitable learning algorithm.

I also plan to use the method of Iyer et al. (2019) to construct smooth SFHs through gaussian processes, and use ML to predict the fractional formation times, rather than the SFR in fixed bins. This will not only remove the need to specify a bin configuration up-front, but also allow greater flexibility in the time resolution of the SFH, allowing for the accommodation of bursts or quiescent periods on arbitrary timescales.

Since the Machine essentially learns a 'prior' of the SFH shape from the simulation, one can use it to learn priors from different simulations, and use these in bayesian SED fitting approaches. One way of extracting priors would be to use increasingly noisy photometry, and show the predictions in the limit of low signal-to-noise.

Another interesting future extension would be to build a model to predict resolved SFHs. This could then be applied to the publicly available SDSS-IV MaNGA catalogues (Goddard et al., 2017). This would rely on simulations having well resolved spatial properties, which is not clearly the case for EAGLE and Illustris; this would then necessitate using higher resolution simulations, which would in turn limit the size of the training set. Future and ongoing simulations, such as the 50 Mpc high resolution Illustris-TNG run, provide a good trade-off between resolution and volume.

7 Further Acknowledgements

For Chapter 3, the authors would like to thank the anonymous referee for useful comments, Daniel Cunnama, Romeel Davé and Kate Storey-Fisher for encouraging discussions, and Yi-Kuan Chiang and Roderik Overzier for their helpful correspondence clarifying the overdensity measurement procedure in Chiang et al. (2013). The Millennium Simulation was carried out by the Virgo Supercomputing Consortium at the Computing Centre of the Max-Planck Society in Garching. The halo merger trees used are publicly available through the German Astronomical Virtual Observatory (GAVO) interface.[29]

For Chapter 4, I wish to thank Paola Santini for helpful discussions related to Santini et al. (2017).

For Chapter 5, the authors wish to thank Rita Tojeiro for help understanding Vespa. VA, KI & EG acknowledge that support for program number HST-AR-14564.001-A and GO-12060 was provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. The Flatiron Institute is supported by the Simons Foundation.

I wish to acknowledge the use of the following open source software packages used throughout this thesis: Scipy (Jones et al., 2001), Astropy (Astropy Collaboration et al., 2013) and Jupyter (Pérez & Granger, 2007). I acknowledge the support of a PhD studentship from the Science and Technology Facilities Council (STFC, grant number ST/P000252/1). PAT acknowledges support from the STFC (grant number ST/P000252/1).

[29] Available at http://www.mpa-garching.mpg.de/millennium/

References

Abazajian K. N., et al., 2009, http://dx.doi.org/10.1088/0067-0049/182/2/543 The Astrophysical Journal Supplement Series, 182, 543

Acquaviva V., Gawiser E., Guaita L., 2011, http://dx.doi.org/10.1088/0004-637X/737/2/47 ApJ, 737, 47

Acquaviva V., Gawiser E., Guaita L., 2012, http://dx.doi.org/10.1017/S1743921312008691 The Spectral Energy Distribution of Galaxies - SED 2011, 284, 42

Acquaviva V., Raichoor A., Gawiser E., 2015, http://dx.doi.org/10.1088/0004-637X/804/1/8 The Astrophysical Journal, 804, 8

Adams J. J., et al., 2011, http://dx.doi.org/10.1088/0067-0049/192/1/5 ApJS, 192, 5

Adams S. M., Martini P., Croxall K. V., Overzier R. A., Silverman J. D., 2015, http://dx.doi.org/10.1093/mnras/stv065 MNRAS, 448, 1335

Allende Prieto C., Lambert D. L., Asplund M., 2001, http://dx.doi.org/10.1086/322874 The Astrophysical Journal Letters, 556, L63

Anders P., Fritze-v. Alvensleben U., 2003, http://dx.doi.org/10.1051/0004-6361:20030151 Astronomy and Astrophysics, 401, 1063

Angulo R. E., White S. D. M., 2010, http://dx.doi.org/10.1111/j.1365-2966.2010.16459.x MNRAS, 405, 143

Astropy Collaboration et al., 2013, http://dx.doi.org/10.1051/0004-6361/201322068 , http://adsabs.harvard.edu/abs/2013A

Bahé Y. M., et al., 2017, http://dx.doi.org/10.1093/mnras/stx1403 Monthly Notices of the Royal Astronomical Society, 470, 4186

Bahé Y. M., et al., 2019, http://dx.doi.org/10.1093/mnras/stz361 Monthly Notices of the Royal Astronomical Society, 485, 2287

Baldry I. K., Glazebrook K., Driver S. P., 2008, http://dx.doi.org/10.1111/j.1365-2966.2008.13348.x Monthly Notices of the Royal Astronomical Society, 388, 945

Ball N. M., Brunner R. J., 2010, http://dx.doi.org/10.1142/S0218271810017160 International Journal of Modern Physics D, 19, 1049

Barbary K., 2016a, extinction v0.3.0, http://dx.doi.org/10.5281/zenodo.804967 doi:10.5281/zenodo.804967, https://zenodo.org/record/804967/export/hx

Barbary K., 2016b, extinction v0.3.0, http://dx.doi.org/10.5281/zenodo.804967 doi:10.5281/zenodo.804967, https://doi.org/10.5281/zenodo.804967

Barnes J., Hut P., 1986, http://dx.doi.org/10.1038/324446a0 Nature, 324, 446

Barnes D. J., Kay S. T., Henson M. A., McCarthy I. G., Schaye J., Jenkins A., 2017a, http://dx.doi.org/10.1093/mnras/stw2722 Mon Not R Astron Soc, 465, 213

Barnes D. J., et al., 2017b, http://dx.doi.org/10.1093/mnras/stx1647 Monthly Notices of the Royal Astronomical Society, 471, 1088

Baron D., 2019, arXiv e-prints, 1904, arXiv:1904.07248

Baugh C. M., 2006, http://dx.doi.org/10.1088/0034-4885/69/12/R02 Reports on Progress in Physics, 69, 3101

Becker R. H., et al., 2001, http://dx.doi.org/10.1086/324231 The Astronomical Journal, 122, 2850

Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013a, http://dx.doi.org/10.1088/0004-637X/762/2/109 The Astrophysical Journal, 762, 109

Behroozi P. S., Wechsler R. H., Conroy C., 2013b, http://dx.doi.org/10.1088/0004-637X/770/1/57 The Astrophysical Journal, 770, 57

Benson A. J., 2010, http://dx.doi.org/10.1016/j.physrep.2010.06.001 Physics Reports, 495, 33

Bergstra J. S., Bardenet R., Bengio Y., Kégl B., 2011, in Advances in neural information processing systems. pp 2546–2554

Bernstein R. A., Freedman W. L., Madore B. F., 2002,] 10.1086/339422, 571, 56

Bett P., Eke V., Frenk C. S., Jenkins A., Helly J., Navarro J., 2007, http://dx.doi.org/10.1111/j.1365-2966.2007.11432.x MNRAS, 376, 215

Bhattacharyya A., 1946, Sankhya: The Indian Journal of Statistics (1933-1960), 7, 401

Bondi H., Hoyle F., 1944, http://dx.doi.org/10.1093/mnras/104.5.273 Monthly Notices of the Royal Astronomical Society, 104, 273

Booth C. M., Schaye J., 2009, http://dx.doi.org/10.1111/j.1365-2966.2009.15043.x Monthly Notices of the Royal Astronomical Society, 398, 53

Borrow J., Bower R. G., Draper P. W., Gonnet P., Schaller M., 2018, arXiv e-prints, p. arXiv:1807.01341

Bouwens R. J., et al., 2012, http://dx.doi.org/10.1088/2041-8205/752/1/L5 ApJL, 752, L5

Bouwens R. J., et al., 2015, http://dx.doi.org/10.1088/0004-637X/803/1/34 The Astrophysical Journal, 803, 34

Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, http://dx.doi.org/10.1111/j.1365-2966.2004.07881.x Monthly Notices of the Royal Astronomical Society, 351, 1151

Bromm V., 2013, http://dx.doi.org/10.1088/0034-4885/76/11/112901 Rep. Prog. Phys., 76, 112901

Bromm V., Larson R. B., 2004, http://dx.doi.org/10.1146/annurev.astro.42.053102.134034 Annual Review of Astronomy and Astrophysics, 42, 79

Bromm V., Yoshida N., 2011, http://dx.doi.org/10.1146/annurev-astro-081710-102608 Annual Review of Astronomy and Astrophysics, 49, 373

Bruzual G., Charlot S., 2003, http://dx.doi.org/10.1046/j.1365-8711.2003.06897.x Monthly Notices of the Royal Astronomical Society, 344, 1000

Bundy K., et al., 2015, http://dx.doi.org/10.1088/0004-637X/798/1/7 The Astrophysical Journal, 798, 7

Byler N., Dalcanton J. J., Conroy C., Johnson B. D., 2017, http://dx.doi.org/10.3847/1538-4357/aa6c66 The Astrophysical Journal, 840, 44

Cai Z., et al., 2016, http://dx.doi.org/10.3847/1538-4357/833/2/135 The Astrophysical Journal, 833, 135

Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, http://dx.doi.org/10.1086/308692 The Astrophysical Journal, 533, 682

Camps P., Trayford J. W., Baes M., Theuns T., Schaller M., Schaye J., 2016, http://dx.doi.org/10.1093/mnras/stw1735 Monthly Notices of the Royal Astronomical Society, 462, 1057

Capak P. L., et al., 2011, http://dx.doi.org/10.1038/nature09681 Nature, 470, 233

Carnall A. C., 2017, preprint, 1705, arXiv:1705.05165

Carnall A. C., Leja J., Johnson B. D., McLure R. J., Dunlop J. S., Conroy C., 2019, http://dx.doi.org/10.3847/1538-4357/ab04a2 The Astrophysical Journal, 873, 44

Carr B. J., Hawking S. W., 1974, http://dx.doi.org/10.1093/mnras/168.2.399 Monthly Notices of the Royal Astronomical Society, 168, 399

Casey C. M., Narayanan D., Cooray A., 2014, http://dx.doi.org/10.1016/j.physrep.2014.02.009 Phys. Rep., 541, 45

Chabrier G., 2003, http://dx.doi.org/10.1086/376392 Publications of the Astronomical Society of the Pacific, 115, 763

Chandrasekhar S., 1931, http://dx.doi.org/10.1086/143324 The Astrophysical Journal, 74, 81

Chevallard J., Charlot S., 2016, http://dx.doi.org/10.1093/mnras/stw1756 Mon Not R Astron Soc, 462, 1415

Chiang Y.-K., Overzier R., Gebhardt K., 2013, http://dx.doi.org/10.1088/0004-637X/779/2/127 ApJ, 779, 127

Chiang Y.-K., Overzier R., Gebhardt K., 2014, http://dx.doi.org/10.1088/2041-8205/782/1/L3 ApJL, 782, L3

Chiang Y.-K., Overzier R. A., Gebhardt K., Henriques B., 2017, http://dx.doi.org/10.3847/2041-8213/aa7e7b ApJL, 844, L23

Chollet F., et al., 2015, Keras, https://keras.io

Cibinel A., et al., 2019, http://dx.doi.org/10.1093/mnras/stz690 Monthly Notices of the Royal Astronomical Society, 485, 5631

Ciesla L., Elbaz D., Fensch J., 2017, http://dx.doi.org/10.1051/0004-6361/201731036 Astronomy and Astrophysics, 608, A41

Clay S., Thomas P., Wilkins S., Henriques B., 2015, http://dx.doi.org/10.1093/mnras/stv818 MNRAS, 451, 2692

Coc A., Vangioni E., 2017, http://dx.doi.org/10.1142/S0218301317410026 International Journal of Modern Physics E, 26, 1741002

Cohn J. D., 2018, http://dx.doi.org/10.1093/mnras/sty1148 Monthly Notices of the Royal Astronomical Society, 478, 2291

Cohn J. D., Battaglia N., 2019, arXiv e-prints, p. arXiv:1905.09920

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, http://dx.doi.org/10.1046/j.1365-8711.2000.03879.x Monthly Notices of the Royal Astronomical Society, 319, 168

Collaboration P., et al., 2018, arXiv e-prints, p. arXiv:1807.06209

Conroy C., 2013, http://dx.doi.org/10.1146/annurev-astro-082812-141017 Annual Review of Astronomy and Astrophysics, 51, 393

Conroy C., Gunn J. E., 2010, http://dx.doi.org/10.1088/0004-637X/712/2/833 The Astrophysical Journal, 712, 833

Conroy C., Gunn J. E., White M., 2009, http://dx.doi.org/10.1088/0004-637X/699/1/486 The Astrophysical Journal, 699, 486

Contini E., Lucia G. D., Hatch N., Borgani S., Kang X., 2016, http://dx.doi.org/10.1093/mnras/stv2852 MNRAS, 456, 1924

Coogan R. T., et al., 2018, http://dx.doi.org/10.1093/mnras/sty1446 Monthly Notices of the Royal Astronomical Society, 479, 703

Cooke E. A., Hatch N. A., Muldrew S. I., Rigby E. E., Kurk J. D., 2014, http://dx.doi.org/10.1093/mnras/stu522 MNRAS, 440, 3262

Cooke E. A., et al., 2015, http://dx.doi.org/10.1093/mnras/stv1413 MNRAS, 452, 2318

Cooke E. A., et al., 2016, http://dx.doi.org/10.3847/0004-637X/816/2/83 ApJ, 816, 83

Crain R. A., et al., 2009, http://dx.doi.org/10.1111/j.1365-2966.2009.15402.x Monthly Notices of the Royal Astronomical Society, 399, 1773

Crain R. A., et al., 2015, http://dx.doi.org/10.1093/mnras/stv725 Monthly Notices of the Royal Astronomical Society, 450, 1937

Croton D. J., et al., 2006, http://dx.doi.org/10.1111/j.1365-2966.2005.09675.x MNRAS, 365, 11

Croton D. J., et al., 2016, http://dx.doi.org/10.3847/0067-0049/222/2/22 ApJS, 222, 22

Daddi E., et al., 2007, http://dx.doi.org/10.1086/521818 ApJ, 670, 156

Dalla Vecchia C., Schaye J., 2012, http://dx.doi.org/10.1111/j.1365-2966.2012.21704.x Monthly Notices of the Royal Astronomical Society, 426, 140

Datta K. K., Ghara R., Majumdar S., Choudhury T. R., Bharadwaj S., Roy H., Datta A., 2016, http://dx.doi.org/10.1007/s12036-016-9405-x Journal of Astrophysics and Astronomy, 37

Davé R., 2008, http://dx.doi.org/10.1111/j.1365-2966.2008.12866.x Mon Not R Astron Soc, 385, 147

Davé R., Thompson R. J., Hopkins P. F., 2016, http://dx.doi.org/10.1093/mnras/stw1862 Monthly Notices of the Royal Astronomical Society, 462, 3265

Davé R., Rafieferantsoa M. H., Thompson R. J., 2017, http://dx.doi.org/10.1093/mnras/stx1693 Monthly Notices of the Royal Astronomical Society, 471, 1671

Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, http://dx.doi.org/10.1093/mnras/stz937 Monthly Notices of the Royal Astronomical Society, 486, 2827

Davidzon I., et al., 2017, http://dx.doi.org/10.1051/0004-6361/201730419 Astronomy and Astrophysics, 605, A70

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, http://dx.doi.org/10.1086/163168 The Astrophysical Journal, 292, 371

Diener C., et al., 2015, http://dx.doi.org/10.1088/0004-637X/802/1/31 ApJ, 802, 31

Dolag K., Borgani S., Murante G., Springel V., 2009, http://dx.doi.org/10.1111/j.1365-2966.2009.15034.x Monthly Notices of the Royal Astronomical Society, 399, 497

Donnari M., et al., 2019, http://dx.doi.org/10.1093/mnras/stz712 Monthly Notices of the Royal Astronomical Society, 485, 4817

Dressler A., 1980, http://dx.doi.org/10.1086/157753 ApJ, 236, 351

Duivenvoorden S., et al., 2016, http://dx.doi.org/10.1093/mnras/stw1466 Monthly Notices of the Royal Astronomical Society, 462, 277

Duncan K., et al., 2014, http://dx.doi.org/10.1093/mnras/stu1622 MNRAS, 444, 2960

Dunlop J. S., Peacock J. A., 1990, Monthly Notices of the Royal Astronomical Society, 247, 19

Edgar R. G., 2004, http://dx.doi.org/10.1016/j.newar.2004.06.001 New Astronomy Reviews, 48, 843

Efstathiou G., Davis M., White S. D. M., Frenk C. S., 1985, http://dx.doi.org/10.1086/191003 The Astrophysical Journal Supplement Series, 57, 241

Elahi P. J., Cañas R., Tobar R. J., Willis J. S., Lagos C. d. P., Power C., Robotham A. S. G., 2019, arXiv e-prints, 1902, arXiv:1902.01010

Eldridge J. J., Stanway E. R., 2016, http://dx.doi.org/10.1093/mnras/stw1772 Monthly Notices of the Royal Astronomical Society, 462, 3302

Eldridge J. J., Stanway E. R., Xiao L., McClelland L. A. S., Taylor G., Ng M., Greis S. M. L., Bray J. C., 2017, http://dx.doi.org/10.1017/pasa.2017.51 Publications of the Astronomical Society of Australia, 34, e058

Ellis R. S., et al., 2013, http://dx.doi.org/10.1088/2041-8205/763/1/L7 The Astrophysical Journal, 763, L7

Ellison S. L., Pettini M., Steidel C. C., Shapley A. E., 2001, http://dx.doi.org/10.1086/319457 ApJ, 549, 770

Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, http://dx.doi.org/10.1093/mnras/stx3298 Monthly Notices of the Royal Astronomical Society, 475, 2978

Fan J., Ma C., Zhong Y., 2019, arXiv e-prints, p. arXiv:1904.05526

Fang J. J., et al., 2018, http://dx.doi.org/10.3847/1538-4357/aabcba The Astrophysical Journal, 858, 100

Fanidakis N., Baugh C. M., Benson A. J., Bower R. G., Cole S., Done C., Frenk C. S., 2011, http://dx.doi.org/10.1111/j.1365-2966.2010.17427.x MNRAS, 410, 53

Farrens S., Abdalla F. B., Cypriano E. S., Sabiu C., Blake C., 2011, http://dx.doi.org/10.1111/j.1365-2966.2011.19356.x Monthly Notices of the Royal Astronomical Society, 417, 1402

Faucher-Giguère C.-A., Lidz A., Zaldarriaga M., Hernquist L., 2009, http://dx.doi.org/10.1088/0004-637X/703/2/1416 The Astrophysical Journal, 703, 1416

Feng Y., Di-Matteo T., Croft R. A., Bird S., Battaglia N., Wilkins S., 2015a, arXiv:1504.06619 [astro-ph]

Feng Y., Di Matteo T., Croft R., Tenneti A., Bird S., Battaglia N., Wilkins S., 2015b, http://dx.doi.org/10.1088/2041-8205/808/1/L17 The Astrophysical Journal, 808, L17

Ferland G. J., et al., 2013, arXiv:1302.4485 [astro-ph]

Ferland G. J., et al., 2017, preprint, 1705, arXiv:1705.10877

Fèvre O. L., Deltorn J. M., Crampton D., Dickinson M., 1996, http://dx.doi.org/10.1086/310319 ApJ, 471, L11

Fèvre O. L., et al., 2015, http://dx.doi.org/10.1051/0004-6361/201423829 A&A, 576, A79

Finkelstein S. L., et al., 2015, http://dx.doi.org/10.1088/0004-637X/810/1/71 The Astrophysical Journal, 810, 71

Fioc M., Rocca-Volmerange B., 1997, Astronomy and Astrophysics, 326, 950

Fioc M., Rocca-Volmerange B., 1999, arXiv e-prints, pp astro–ph/9912179

Fioc M., Rocca-Volmerange B., 2019, arXiv e-prints, 1902, arXiv:1902.02198

Fitzpatrick E. L., 1999, http://dx.doi.org/10.1086/316293 Publications of the Astronomical Society of the Pacific, 111, 63

Flamary R., 2016, http://dx.doi.org/10.23919/eusipco.2017.8081654 2017 25th European Signal Processing Conference (EUSIPCO), pp 2468–2472

Foreman-Mackey D., Sick J., Johnson B., 2014, python-fsps: Python bindings to FSPS (v0.1.1), http://dx.doi.org/10.5281/zenodo.12157 doi:10.5281/zenodo.12157, https://doi.org/10.5281/zenodo.12157

Franck J. R., McGaugh S. S., 2016a, http://dx.doi.org/10.3847/0004-637X/817/2/158 ApJ, 817, 158

Franck J. R., McGaugh S. S., 2016b, http://dx.doi.org/10.3847/0004-637X/833/1/15 The Astrophysical Journal, 833, 15

Franx M., Illingworth G., de Zeeuw T., 1991, http://dx.doi.org/10.1086/170769 ApJ, 383, 112

Friedman A., 1922, http://dx.doi.org/10.1007/BF01332580 Z. Physik, 10, 377

Furlong M., et al., 2015, http://dx.doi.org/10.1093/mnras/stv852 Monthly Notices of the Royal Astronomical Society, 450, 4486

Galametz A., et al., 2010, http://dx.doi.org/10.1051/0004-6361/201015035 A&A, 522, A58

Genel S., et al., 2014, http://dx.doi.org/10.1093/mnras/stu1654 Monthly Notices of the Royal Astronomical Society, 445, 175

Geurts P., Ernst D., Wehenkel L., 2006, http://dx.doi.org/10.1007/s10994-006-6226-1 Mach Learn, 63, 3

Glover S. C. O., 2013, http://dx.doi.org/10.1007/978-3-642-32362-1_3 arXiv:1209.2509, 396, 103

Goddard D., et al., 2017, http://dx.doi.org/10.1093/mnras/stw2719 Monthly Notices of the Royal Astronomical Society, 465, 688

González Delgado R. M., et al., 2017, http://dx.doi.org/10.1051/0004-6361/201730883 Astronomy and Astrophysics, 607, A128

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, http://dx.doi.org/10.1093/mnras/stt2410 Monthly Notices of the Royal Astronomical Society, 439, 264

Greif T. H., 2014, arXiv:1410.3482 [astro-ph]

Greif T. H., Springel V., White S. D. M., Glover S. C. O., Clark P. C., Smith R. J., Klessen R. S., Bromm V., 2011, http://dx.doi.org/10.1088/0004-637X/737/2/75 ApJ, 737, 75

Gu J., et al., 2018, http://dx.doi.org/10.1016/j.patcog.2017.10.013 Pattern Recognition, 77, 354

Gunawardhana M. L. P., et al., 2011, http://dx.doi.org/10.1111/j.1365-2966.2011.18800.x Monthly Notices of the Royal Astronomical Society, 415, 1647

Gunn J. E., Peterson B. A., 1965, http://dx.doi.org/10.1086/148444 The Astrophysical Journal, 142, 1633

Guo Q., et al., 2011, http://dx.doi.org/10.1111/j.1365-2966.2010.18114.x MNRAS, 413, 101

Guth A. H., 1981, http://dx.doi.org/10.1103/PhysRevD.23.347 Phys. Rev. D, 23, 347

Haardt F., Madau P., 2012, http://dx.doi.org/10.1088/0004-637X/746/2/125 ApJ, 746, 125

Habouzit M., Volonteri M., Somerville R. S., Dubois Y., Peirani S., Pichon C., Devriendt J., 2018, arXiv e-prints, p. arXiv:1810.11535

Harikane Y., et al., 2019, arXiv e-prints, 1902, arXiv:1902.09555

Hassan S., Davé R., Mitra S., Finlator K., Ciardi B., Santos M. G., 2017, arXiv:1705.05398 [astro-ph]

Hatch N. A., et al., 2011a, http://dx.doi.org/10.1111/j.1365-2966.2010.17538.x MNRAS, 410, 1537

Hatch N. A., Kurk J. D., Pentericci L., Venemans B. P., Kuiper E., Miley G. K., Röttgering H. J. A., 2011b, http://dx.doi.org/10.1111/j.1365-2966.2011.18735.x MNRAS, 415, 2993

Hatch N. A., et al., 2014, http://dx.doi.org/10.1093/mnras/stu1725 MNRAS, 445, 280

Hayashi M., Kodama T., Tanaka I., Shimakawa R., Koyama Y., Tadaki K.-i., Suzuki T. L., Yamamoto M., 2016, http://dx.doi.org/10.3847/2041-8205/826/2/L28 The Astrophysical Journal Letters, 826, L28

Heavens A. F., Jimenez R., Lahav O., 2000, http://dx.doi.org/10.1046/j.1365-8711.2000.03692.x Monthly Notices of the Royal Astronomical Society, 317, 965

Hennawi J. F., Prochaska J. X., Cantalupo S., Arrigoni-Battaia F., 2015, http://dx.doi.org/10.1126/science.aaa5397 Science, 348, 779

Henriques B. M. B., Thomas P. A., Oliver S., Roseboom I., 2009, http://dx.doi.org/10.1111/j.1365-2966.2009.14730.x Monthly Notices of the Royal Astronomical Society, 396, 535

Henriques B. M. B., White S. D. M., Thomas P. A., Angulo R., Guo Q., Lemson G., Springel V., Overzier R., 2015, http://dx.doi.org/10.1093/mnras/stv705 MNRAS, 451, 2663

Hernquist L., Bouchet F. R., Suto Y., 1991, http://dx.doi.org/10.1086/191530 The Astrophysical Journal Supplement Series, 75, 231

Higuchi R., et al., 2018, arXiv:1801.00531 [astro-ph]

Hinshaw G., et al., 2013, http://dx.doi.org/10.1088/0067-0049/208/2/19 The Astrophysical Journal Supplement Series, 208, 19

Hockney R. W., Eastwood J. W., 1988, Computer simulation using particles. http://adsabs.harvard.edu/abs/1988csup.book.....H

Hopkins A. M., 2018, http://dx.doi.org/10.1017/pasa.2018.29 Publications of the Astronomical Society of Australia, 35

Hopkins P. F., Richards G. T., Hernquist L., 2007, http://dx.doi.org/10.1086/509629 ApJ, 654, 731

Husband K., Bremer M. N., Stanway E. R., Davies L. J. M., Lehnert M. D., Douglas L. S., 2013, http://dx.doi.org/10.1093/mnras/stt642 MNRAS, 432, 2869

Ilbert O., et al., 2013, http://dx.doi.org/10.1051/0004-6361/201321100 Astronomy and Astrophysics, 556, A55

Iliev I. T., et al., 2006, http://dx.doi.org/10.1111/j.1365-2966.2006.10775.x Mon Not R Astron Soc, 371, 1057

Iliev I. T., et al., 2009, http://dx.doi.org/10.1111/j.1365-2966.2009.15558.x Mon Not R Astron Soc, 400, 1283

Iyer K., Gawiser E., 2017, http://dx.doi.org/10.3847/1538-4357/aa63f0 ApJ, 838, 127

Iyer K. G., Gawiser E., Faber S. M., Ferguson H. C., Kartaltepe J., Koekemoer A. M., Pacifici C., Somerville R. S., 2019, http://dx.doi.org/10.3847/1538-4357/ab2052 The Astrophysical Journal, 879, 116

Jarvis M. J., Rawlings S., Willott C. J., Blundell K. M., Eales S., Lacy M., 2001, http://dx.doi.org/10.1046/j.1365-8711.2001.04778.x Monthly Notices of the Royal Astronomical Society, 327, 907

Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, http://dx.doi.org/10.1093/mnras/stu390 Mon Not R Astron Soc, 440, 2115

Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/

Kaiser N., 1987, http://dx.doi.org/10.1093/mnras/227.1.1 Monthly Notices of the Royal Astronomical Society, 227, 1

Kajisawa M., Kodama T., Tanaka I., Yamada T., Bower R., 2006, http://dx.doi.org/10.1111/j.1365-2966.2006.10704.x Mon Not R Astron Soc, 371, 577

Kamdar H. M., Turk M. J., Brunner R. J., 2016, http://dx.doi.org/10.1093/mnras/stv2981 MNRAS, 457, 1162

Katsianis A., Tescari E., Wyithe J. S. B., 2016, http://dx.doi.org/10.1017/pasa.2016.21 Publications of the Astronomical Society of Australia, 33

Katsianis A., et al., 2017, http://dx.doi.org/10.1093/mnras/stx2020 Monthly Notices of the Royal Astronomical Society, 472, 919

Katsianis A., et al., 2019, arXiv:1905.02023 [astro-ph]

Katz N., White S. D. M., 1993, http://dx.doi.org/10.1086/172935 The Astrophysical Journal, 412, 455

Kauffmann G., White S. D. M., Guiderdoni B., 1993, http://dx.doi.org/10.1093/mnras/264.1.201 Monthly Notices of the Royal Astronomical Society, 264, 201

Kennicutt R. C., 1998, http://dx.doi.org/10.1146/annurev.astro.36.1.189 Annu. Rev. Astron. Astrophys., 36, 189

Kennicutt Jr R. C., Evans II N. J., 2012, http://dx.doi.org/10.1146/annurev-astro-081811-125610 Annu. Rev. Astron. Astrophys., 50, 531

Kingma D. P., Ba J., 2014, preprint, 1412, arXiv:1412.6980

Kiranyaz S., Avci O., Abdeljaber O., Ince T., Gabbouj M., Inman D. J., 2019, arXiv e-prints, p. arXiv:1905.03554

Klypin A. A., Trujillo-Gomez S., Primack J., 2011, http://dx.doi.org/10.1088/0004-637X/740/2/102 The Astrophysical Journal, 740, 102

Knebe A., et al., 2011, http://dx.doi.org/10.1111/j.1365-2966.2011.18858.x Monthly Notices of the Royal Astronomical Society, 415, 2293

Knebe A., et al., 2013, http://dx.doi.org/10.1093/mnras/stt1403 Monthly Notices of the Royal Astronomical Society, 435, 1618

Koyama Y., Kodama T., Tadaki K.-i., Hayashi M., Tanaka M., Smail I., Tanaka I., Kurk J., 2012, http://dx.doi.org/10.1093/mnras/sts133 MNRAS

Koyama Y., et al., 2013, http://dx.doi.org/10.1093/mnras/stt1035 MNRAS, 434, 423

Kroupa P., 2001, http://dx.doi.org/10.1046/j.1365-8711.2001.04022.x Monthly Notices of the Royal Astronomical Society, 322, 231

Kroupa P., Weidner C., Pflamm-Altenburg J., Thies I., Dabringhausen J., Marks M., Maschberger T., 2013, in Oswalt T. D., Gilmore G., eds, , Planets, Stars and Stellar Systems: Volume 5: Galactic Structure and Stellar Populations. Springer Netherlands, Dordrecht, pp 115–242, http://dx.doi.org/10.1007/978-94-007-5612-0_4 doi:10.1007/978-94-007-5612-0_4, https://doi.org/10.1007/978-94-007-5612-0_4

Lagos C. d. P., Tobar R. J., Robotham A. S. G., Obreschkow D., Mitchell P. D., Power C., Elahi P. J., 2018, preprint, 1807, arXiv:1807.11180

Larson R. B., 1998, http://dx.doi.org/10.1046/j.1365-8711.1998.02045.x Monthly Notices of the Royal Astronomical Society, 301, 569

Lee-Brown D. B., et al., 2017, http://dx.doi.org/10.3847/1538-4357/aa7948 The Astrophysical Journal, 844, 43

Lee J. C., et al., 2009, http://dx.doi.org/10.1088/0004-637X/706/1/599 The Astrophysical Journal, 706, 599

Lee N., et al., 2015, http://dx.doi.org/10.1088/0004-637X/801/2/80 ApJ, 801, 80

Lee M. M., et al., 2017, http://dx.doi.org/10.3847/1538-4357/aa74c2 The Astrophysical Journal, 842, 55

Leja J., Dokkum P. G. v., Franx M., Whitaker K. E., 2015, http://dx.doi.org/10.1088/0004-637X/798/2/115 ApJ, 798, 115

Leja J., Johnson B. D., Conroy C., van Dokkum P. G., Byler N., 2017, http://dx.doi.org/10.3847/1538-4357/aa5ffe The Astrophysical Journal, 837, 170

Leja J., et al., 2018, arXiv e-prints, 1812, arXiv:1812.05608

Leja J., Carnall A. C., Johnson B. D., Conroy C., Speagle J. S., 2019, http://dx.doi.org/10.3847/1538-4357/ab133c The Astrophysical Journal, 876, 3

Lemaux B. C., et al., 2017, arXiv:1703.10170 [astro-ph]

Liddle A. R., Lyth D. H., 2000, Cosmological Inflation and Large-Scale Structure. http://adsabs.harvard.edu/abs/2000cils.book.....L

Lovell C. C., Thomas P. A., Wilkins S. M., 2018, http://dx.doi.org/10.1093/mnras/stx3090 Monthly Notices of the Royal Astronomical Society, 474, 4612

Lu Y., Mo H. J., Weinberg M. D., Katz N., 2011, http://dx.doi.org/10.1111/j.1365-2966.2011.19170.x Mon Not R Astron Soc, 416, 1949

Lucia G. D., Blaizot J., 2007, http://dx.doi.org/10.1111/j.1365-2966.2006.11287.x MNRAS, 375, 2

Ma X., Hopkins P. F., Kasen D., Quataert E., Faucher-Giguere C.-A., Keres D., Murray N., 2016, arXiv:1601.07559 [astro-ph]

Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9, 2579

Madau P., Dickinson M., 2014, http://dx.doi.org/10.1146/annurev-astro-081811-125615 Annual Review of Astronomy and Astrophysics, 52, 415

Madau P., Haardt F., 2015, http://dx.doi.org/10.1088/2041-8205/813/1/L8 The Astrophysical Journal, 813, L8

Magorrian J., et al., 1998, http://dx.doi.org/10.1086/300353 The Astronomical Journal, 115, 2285

Maraston C., 2005, http://dx.doi.org/10.1111/j.1365-2966.2005.09270.x Monthly Notices of the Royal Astronomical Society, 362, 799

Maraston C., Pforr J., Renzini A., Daddi E., Dickinson M., Cimatti A., Tonini C., 2010, http://dx.doi.org/10.1111/j.1365-2966.2010.16973.x Monthly Notices of the Royal Astronomical Society, 407, 830

Markevitch M., Gonzalez A. H., Clowe D., Vikhlinin A., Forman W., Jones C., Murray S., Tucker W., 2004, http://dx.doi.org/10.1086/383178 ApJ, 606, 819

Masters D., et al., 2015, http://dx.doi.org/10.1088/0004-637X/813/1/53 The Astrophysical Journal, 813, 53

Matsuda Y., et al., 2005, http://dx.doi.org/10.1086/499071 ApJ, 634, L125

Matthee J., Schaye J., 2019, http://dx.doi.org/10.1093/mnras/stz030 Monthly Notices of the Royal Astronomical Society, 484, 915

Matthee J., Schaye J., Crain R. A., Schaller M., Bower R., Theuns T., 2017, http://dx.doi.org/10.1093/mnras/stw2884 Monthly Notices of the Royal Astronomical Society, 465, 2381

Mazzucchelli C., Bañados E., Decarli R., Farina E. P., Venemans B. P., Walter F., Overzier R., 2017, http://dx.doi.org/10.3847/1538-4357/834/1/83 The Astrophysical Journal, 834, 83

McAlpine S., et al., 2016, http://dx.doi.org/10.1016/j.ascom.2016.02.004 Astronomy and Computing, 15, 72

Mellema G., et al., 2013, http://dx.doi.org/10.1007/s10686-013-9334-5 Experimental Astronomy, 36, 235

Miley G. K., et al., 2006, http://dx.doi.org/10.1086/508534 ApJ, 650, L29

Miller G. E., Scalo J. M., 1979, http://dx.doi.org/10.1086/190629 The Astrophysical Journal Supplement Series, 41, 513

Miller T. B., Chapman S. C., Hayward C. C., Behroozi P. S., Bradford C. M., Willott C. J., Wagg J., 2016, arXiv:1611.08552 [astro-ph]

Mobasher B., et al., 2015, http://dx.doi.org/10.1088/0004-637X/808/1/101 The Astrophysical Journal, 808, 101

Møller P., Fynbo J. U., 2001, http://dx.doi.org/10.1051/0004-6361:20010606 A&A, 372

Monaco P., Møller P., Fynbo J. P. U., Weidinger M., Ledoux C., Theuns T., 2005, http://dx.doi.org/10.1051/0004-6361:20042570 A&A, 440

Morselli L., et al., 2014, http://dx.doi.org/10.1051/0004-6361/201423853 A&A, 568, A1

Moustakas J., et al., 2013, http://dx.doi.org/10.1088/0004-637X/767/1/50 The Astrophysical Journal, 767, 50

Muldrew S. I., et al., 2012, http://dx.doi.org/10.1111/j.1365-2966.2011.19922.x MNRAS, 419, 2670

Muldrew S. I., Hatch N. A., Cooke E. A., 2015, http://dx.doi.org/10.1093/mnras/stv1449 MNRAS, 452, 2528

Muldrew S. I., Hatch N. A., Cooke E. A., 2018, http://dx.doi.org/10.1093/mnras/stx2454 Monthly Notices of the Royal Astronomical Society, 473, 2335

Muzzin A., et al., 2013, http://dx.doi.org/10.1088/0004-637X/777/1/18 ApJ, 777, 18

Narayanan D., Dave R., Johnson B., Thompson R., Conroy C., Geach J. E., 2017, arXiv:1705.05858 [astro-ph]

Neistein E., van den Bosch F. C., Dekel A., 2006, http://dx.doi.org/10.1111/j.1365-2966.2006.10918.x Monthly Notices of the Royal Astronomical Society, 372, 933

Newman A. B., Ellis R. S., Andreon S., Treu T., Raichoor A., Trinchieri G., 2014, http://dx.doi.org/10.1088/0004-637X/788/1/51 The Astrophysical Journal, 788, 51

Noeske K. G., et al., 2007, http://dx.doi.org/10.1086/517926 The Astrophysical Journal Letters, 660, L43

O'Donnell J. E., 1994, http://dx.doi.org/10.1086/173713 The Astrophysical Journal, 422, 158

Oppenheimer B. D., Davé R., 2008, http://dx.doi.org/10.1111/j.1365-2966.2008.13280.x Monthly Notices of the Royal Astronomical Society, 387, 577

Orsi Á. A., Fanidakis N., Lacey C. G., Baugh C. M., 2016, http://dx.doi.org/10.1093/mnras/stv2919 MNRAS, 456, 3827

Ouchi M., et al., 2005, http://dx.doi.org/10.1086/428499 ApJ, 620, L1

Overzier R. A., 2016, http://dx.doi.org/10.1007/s00159-016-0100-3 Astronomy and Astrophysics Review, 24, 14

Overzier R. A., Guo Q., Kauffmann G., Lucia G. D., Bouwens R., Lemson G., 2009, http://dx.doi.org/10.1111/j.1365-2966.2008.14264.x MNRAS, 394, 577

Pacifici C., Kassin S. A., Weiner B., Charlot S., Gardner J. P., 2013, http://dx.doi.org/10.1088/2041-8205/762/1/L15 The Astrophysical Journal, 762, L15

Pacifici C., et al., 2014, http://dx.doi.org/10.1093/mnras/stu2447 Monthly Notices of the Royal Astronomical Society, 447, 786

Parsa S., Dunlop J. S., McLure R. J., 2017, preprint, 1704, arXiv:1704.07750

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peebles P. J. E., 1984, http://dx.doi.org/10.1086/162425 The Astrophysical Journal, 284, 439

Peebles P. J. E., 1993, Principles of Physical Cosmology. http://adsabs.harvard.edu/abs/1993ppc..book.....P

Peng Y.-j., et al., 2010, http://dx.doi.org/10.1088/0004-637X/721/1/193 The Astrophysical Journal, 721, 193

Pentericci L., et al., 2000, Astronomy and Astrophysics, 361, L25

Penzias A. A., Wilson R. W., 1965, http://dx.doi.org/10.1086/148307 The Astrophysical Journal, 142, 419

Pérez F., Granger B. E., 2007, http://dx.doi.org/10.1109/MCSE.2007.53 Comput. Sci. Eng., 9, 21

Petrillo C. E., et al., 2017, http://dx.doi.org/10.1093/mnras/stx2052 Monthly Notices of the Royal Astronomical Society, 472, 1129

Pillepich A., et al., 2017, arXiv:1707.03406 [astro-ph]

Pirzkal N., Rothberg B., Nilsson K. K., Finkelstein S., Koekemoer A., Malhotra S., Rhoads J., 2012, http://dx.doi.org/10.1088/0004-637X/748/2/122 ApJ, 748, 122

Planck Collaboration et al., 2014, http://dx.doi.org/10.1051/0004-6361/201321529 A&A, 571, A1

Poole G. B., Angel P. W., Mutch S. J., Power C., Duffy A. R., Geil P. M., Mesinger A., Wyithe S. B., 2016, http://dx.doi.org/10.1093/mnras/stw674 Monthly Notices of the Royal Astronomical Society, 459, 3025

Price S. H., et al., 2014, http://dx.doi.org/10.1088/0004-637X/788/1/86 ApJ, 788, 86

Qu Y., et al., 2016, arXiv:1609.07243 [astro-ph]

Ragone-Figueroa C., Granato G. L., Ferraro M. E., Murante G., Biffi V., Borgani S., Planelles S., Rasia E., 2018, http://dx.doi.org/10.1093/mnras/sty1639 Mon Not R Astron Soc, 479, 1125

Rahmati A., Pawlik A. H., Raičevic? M., Schaye J., 2013, http://dx.doi.org/10.1093/mnras/stt066 MNRAS, 430, 2427

Ramos Almeida C., Bessiere P. S., Tadhunter C. N., Inskip K. J., Morganti R., Dicken D., González-Serrano J. I., Holt J., 2013, http://dx.doi.org/10.1093/mnras/stt1595 MNRAS, 436, 997

Reddy N. A., et al., 2015, http://dx.doi.org/10.1088/0004-637X/806/2/259 ApJ, 806, 259

Rees M. J., Ostriker J. P., 1977, http://dx.doi.org/10.1093/mnras/179.4.541 Monthly Notices of the Royal Astronomical Society, 179, 541

Reines A. E., Nidever D. L., Whelan D. G., Johnson K. E., 2010, http://dx.doi.org/10.1088/0004-637X/708/1/26 The Astrophysical Journal, 708, 26

Rigby E. E., Best P. N., Brookes M. H., Peacock J. A., Dunlop J. S., Röttgering H. J. A., Wall J. V., Ker L., 2011, http://dx.doi.org/10.1111/j.1365-2966.2011.19167.x Monthly Notices of the Royal Astronomical Society, 416, 1900

Robertson B. E., Ellis R. S., Furlanetto S. R., Dunlop J. S., 2015, http://dx.doi.org/10.1088/2041-8205/802/2/L19 The Astrophysical Journal, 802, L19

Rodriguez-Gomez V., et al., 2015, http://dx.doi.org/10.1093/mnras/stv264 Monthly Notices of the Royal Astronomical Society, 449, 49

Rosas-Guevara Y. M., et al., 2015, http://dx.doi.org/10.1093/mnras/stv2056 Monthly Notices of the Royal Astronomical Society, 454, 1038

Sajina A., Scott D., Dennefeld M., Dole H., Lacy M., Lagache G., 2006, http://dx.doi.org/10.1111/j.1365-2966.2006.10361.x Monthly Notices of the Royal Astronomical Society, 369, 939

Salmon B., et al., 2015, http://dx.doi.org/10.1088/0004-637X/799/2/183 ApJ, 799, 183

Salpeter E. E., 1955, http://dx.doi.org/10.1086/145971 The Astrophysical Journal, 121, 161

Santini P., et al., 2009, http://dx.doi.org/10.1051/0004-6361/200811434 A&A, 504, 751

Santini P., et al., 2017, http://dx.doi.org/10.3847/1538-4357/aa8874 The Astrophysical Journal, 847, 76

Schaller M., Dalla Vecchia C., Schaye J., Bower R. G., Theuns T., Crain R. A., Furlong M., McCarthy I. G., 2015, http://dx.doi.org/10.1093/mnras/stv2169 Mon Not R Astron Soc, 454, 2277

Schaye J., 2004, http://dx.doi.org/10.1086/421232 The Astrophysical Journal, 609, 667

Schaye J., Dalla Vecchia C., 2008, http://dx.doi.org/10.1111/j.1365-2966.2007.12639.x Monthly Notices of the Royal Astronomical Society, 383, 1210

Schaye J., et al., 2014, http://dx.doi.org/10.1093/mnras/stu2058 Monthly Notices of the Royal Astronomical Society, 446, 521

Schechter P., 1976, http://dx.doi.org/10.1086/154079 The Astrophysical Journal, 203, 297

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, http://dx.doi.org/10.1086/305772 The Astrophysical Journal, 500, 525

Schneider M. D., Frenk C. S., Cole S., 2012, http://dx.doi.org/10.1088/1475-7516/2012/05/030 JCAP, 2012, 030

Schreiber C., et al., 2015, http://dx.doi.org/10.1051/0004-6361/201425017 A&A, 575, A74

Shattow G. M., Croton D. J., Skibba R. A., Muldrew S. I., Pearce F. R., Abbas U., 2013, http://dx.doi.org/10.1093/mnras/stt998 MNRAS, 433, 3314

Shi K., et al., 2019, arXiv e-prints, 1905, arXiv:1905.06337

Shimakawa R., Kodama T., Tadaki K.-i., Tanaka I., Hayashi M., Koyama Y., 2014, http://dx.doi.org/10.1093/mnrasl/slu029 Monthly Notices of the Royal Astronomical Society: Letters, 441, L1

Shimakawa R., et al., 2017a, http://dx.doi.org/10.1093/mnras/stx2494 Monthly Notices of the Royal Astronomical Society

Shimakawa R., Koyama Y., Prochaska J. X., Guo Y., Tadaki K.-i., Kodama T., 2017b, arXiv e-prints, p. arXiv:1705.01127

Shimakawa R., et al., 2018, http://dx.doi.org/10.1093/mnras/sty2618 Monthly Notices of the Royal Astronomical Society, 481, 5630

Shimasaku K., et al., 2003, http://dx.doi.org/10.1086/374880 ApJ, 586, L111

Shivaei I., et al., 2015, http://dx.doi.org/10.1088/0004-637X/815/2/98 ApJ, 815, 98

Sijacki D., Springel V., Di Matteo T., Hernquist L., 2007, http://dx.doi.org/10.1111/j.1365-2966.2007.12153.x Monthly Notices of the Royal Astronomical Society, 380, 877

Silk J., Wyse R. F. G., 1993, http://dx.doi.org/10.1016/0370-1573(93)90174-C Physics Reports, 231, 293

Simet M., Chartab Soltani N., Lu Y., Mobasher B., 2019, arXiv e-prints, p. arXiv:1905.08996

Simha V., Weinberg D. H., Conroy C., Dave R., Fardal M., Katz N., Oppenheimer B. D., 2014, preprint, 1404, arXiv:1404.0402

Smith A., Bromm V., 2019, arXiv e-prints, 1904, arXiv:1904.12890

Smith C. M. A., Gear W. K., Smith M. W. L., Papageorgiou A., Eales S. A., 2019, arXiv e-prints, 1904, arXiv:1904.07246

Somerville R. S., Davé R., 2015, http://dx.doi.org/10.1146/annurev-astro-082812-140951 Annual Review of Astronomy and Astrophysics, 53, 51

Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, http://dx.doi.org/10.1111/j.1365-2966.2008.13805.x Monthly Notices of the Royal Astronomical Society, 391, 481

Sousbie T., 2011, http://dx.doi.org/10.1111/j.1365-2966.2011.18394.x MNRAS, 414, 350

Sparre M., et al., 2015, http://dx.doi.org/10.1093/mnras/stu2713 Monthly Notices of the Royal Astronomical Society, 447, 3548

Speagle J. S., Steinhardt C. L., Capak P. L., Silverman J. D., 2014, http://dx.doi.org/10.1088/0067-0049/214/2/15 The Astrophysical Journal Supplement Series, 214, 15

Spergel D. N., et al., 2003, http://dx.doi.org/10.1086/377226 ApJS, 148, 175

Spitler L. R., et al., 2012, http://dx.doi.org/10.1088/2041-8205/748/2/L21 ApJL, 748, L21

Springel V., 2010a, http://dx.doi.org/10.1146/annurev-astro-081309-130914 Annual Review of Astronomy and Astrophysics, 48, 391

Springel V., 2010b, http://dx.doi.org/10.1111/j.1365-2966.2009.15715.x Mon Not R Astron Soc, 401, 791

Springel V., Hernquist L., 2003, http://dx.doi.org/10.1046/j.1365-8711.2003.06206.x Monthly Notices of the Royal Astronomical Society, 339, 289

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, http://dx.doi.org/10.1046/j.1365-8711.2001.04912.x Monthly Notices of the Royal Astronomical Society, 328, 726

Springel V., et al., 2005, http://dx.doi.org/10.1038/nature03597 Nature, 435, 629

Srisawat C., et al., 2013, http://dx.doi.org/10.1093/mnras/stt1545 Monthly Notices of the Royal Astronomical Society, 436, 150

Stanway E. R., Eldridge J. J., 2018, http://dx.doi.org/10.1093/mnras/sty1353 Monthly Notices of the Royal Astronomical Society, 479, 75

Stanway E. R., Eldridge J. J., Becker G. D., 2016, http://dx.doi.org/10.1093/mnras/stv2661 Monthly Notices of the Royal Astronomical Society, 456, 485

Steidel C. C., Adelberger K. L., Dickinson M., Giavalisco M., Pettini M., Kellogg M., 1998, http://dx.doi.org/10.1086/305073 ApJ, 492, 428

Steidel C. C., Adelberger K. L., Shapley A. E., Pettini M., Dickinson M., Giavalisco M., 2000, http://dx.doi.org/10.1086/308568 ApJ, 532, 170

Steidel C. C., Adelberger K. L., Shapley A. E., Erb D. K., Reddy N. A., Pettini M., 2005, http://dx.doi.org/10.1086/429989 ApJ, 626, 44

Strauss M. A., et al., 2002, http://dx.doi.org/10.1086/342343 The Astronomical Journal, 124, 1810

Strazzullo V., et al., 2013, http://dx.doi.org/10.1088/0004-637X/772/2/118 The Astrophysical Journal, 772, 118

Strazzullo V., et al., 2018, http://dx.doi.org/10.3847/1538-4357/aacd10 The Astrophysical Journal, 862, 64

Suwa T., Habe A., Yoshikawa K., 2006, http://dx.doi.org/10.1086/506607 ApJ, 646, L5

Tanaka I., et al., 2011, http://dx.doi.org/10.1093/pasj/63.sp2.S415 Publications of the Astronomical Society of Japan, 63, 415

Tasca L. a. M., et al., 2015, http://dx.doi.org/10.1051/0004-6361/201425379 A&A, 581, A54

Taylor E. N., et al., 2015, http://dx.doi.org/10.1093/mnras/stu1900 Monthly Notices of the Royal Astronomical Society, 446, 2144

Tegmark M., et al., 2004, http://dx.doi.org/10.1103/PhysRevD.69.103501 Phys. Rev. D, 69, 103501

The EAGLE team 2017, preprint, 1706, arXiv:1706.09899

Thomas P. A., et al., 1998, http://dx.doi.org/10.1046/j.1365-8711.1998.01491.x MNRAS, 296, 1061

Tojeiro R., Heavens A. F., Jimenez R., Panter B., 2007, http://dx.doi.org/10.1111/j.1365-2966.2007.12323.x Monthly Notices of the Royal Astronomical Society, 381, 1252

Tojeiro R., Wilkins S., Heavens A. F., Panter B., Jimenez R., 2009, http://dx.doi.org/10.1088/0067-0049/185/1/1 The Astrophysical Journal Supplement Series, 185, 1

Tomczak A. R., et al., 2014, http://dx.doi.org/10.1088/0004-637X/783/2/85 The Astrophysical Journal, 783, 85

Tormen G., Bouchet F. R., White S. D. M., 1997, http://dx.doi.org/10.1093/mnras/286.4.865 Mon Not R Astron Soc, 286, 865

Torrey P., et al., 2015, http://dx.doi.org/10.1093/mnras/stu2592 Monthly Notices of the Royal Astronomical Society, 447, 2753

Toshikawa J., et al., 2012, http://dx.doi.org/10.1088/0004-637X/750/2/137 ApJ, 750, 137

Toshikawa J., et al., 2016, http://dx.doi.org/10.3847/0004-637X/826/2/114 ApJ, 826, 114

Toshikawa J., et al., 2017, preprint, 1708, arXiv:1708.09421

Trayford J. W., et al., 2015, http://dx.doi.org/10.1093/mnras/stv1461 Monthly Notices of the Royal Astronomical Society, 452, 2879

Trayford J. W., et al., 2017, http://dx.doi.org/10.1093/mnras/stx1051 Monthly Notices of the Royal Astronomical Society, 470, 771

Trombetti T., Burigana C., 2018, http://dx.doi.org/10.3389/fspas.2018.00033 Front. Astron. Space Sci., 5

Tuccillo D., Huertas-Company M., Decencière E., Velasco-Forero S., 2017. eprint: arXiv:1701.05917, pp 191–196, http://dx.doi.org/10.1017/S1743921317000552 doi:10.1017/S1743921317000552, http://adsabs.harvard.edu/abs/2017IAUS..325..191T

Uchiyama H., et al., 2017, arXiv:1704.06050 [astro-ph]

Venemans B. P., et al., 2002, http://dx.doi.org/10.1086/340563 ApJ, 569, L11

Venemans B. P., et al., 2004, http://dx.doi.org/10.1051/0004-6361:200400041 Astronomy and Astrophysics, 424, L17

Venemans B. P., et al., 2005, http://dx.doi.org/10.1051/0004-6361:20042038 Astronomy and Astrophysics, 431, 793

Venemans B. P., et al., 2007, http://dx.doi.org/10.1051/0004-6361:20053941 A&A, 461, 823

Vijayan A. P., Clay S. J., Thomas P. A., Yates R. M., Wilkins S. M., Henriques B. M., 2019, arXiv e-prints, 1904, arXiv:1904.02196

Vikhlinin A. A., Kravtsov A. V., Markevich M. L., Sunyaev R. A., Churazov E. M., 2014, http://dx.doi.org/10.3367/UFNe.0184.201404a.0339 Phys.-Usp., 57, 317

Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2013, http://dx.doi.org/10.1093/mnras/stt1789 Monthly Notices of the Royal Astronomical Society, 436, 3031

Vogelsberger M., et al., 2014, http://dx.doi.org/10.1093/mnras/stu1536 Monthly Notices of the Royal Astronomical Society, 444, 1518

Vulcani B., et al., 2011, http://dx.doi.org/10.1111/j.1365-2966.2010.17904.x MNRAS, 412, 246

Walcher J., Groves B., Budavári T., Dale D., 2011, http://dx.doi.org/10.1007/s10509-010-0458-z Astrophysics and Space Science, 331, 1

Wattenberg M., Viégas F., Johnson I., 2016, http://dx.doi.org/10.23915/distill.00002 Distill

Whitaker K. E., et al., 2011, http://dx.doi.org/10.1088/0004-637X/735/2/86 ApJ, 735, 86

Whitaker K. E., et al., 2014, http://dx.doi.org/10.1088/0004-637X/795/2/104 ApJ, 795, 104

White S. D. M., Frenk C. S., 1991, http://dx.doi.org/10.1086/170483 The Astrophysical Journal, 379, 52

White S. D. M., Rees M. J., 1978, http://dx.doi.org/10.1093/mnras/183.3.341 Monthly Notices of the Royal Astronomical Society, 183, 341

Wiersma R. P. C., Schaye J., Smith B. D., 2009a, http://dx.doi.org/10.1111/j.1365-2966.2008.14191.x MNRAS, 393, 99

Wiersma R. P. C., Schaye J., Theuns T., Dalla Vecchia C., Tornatore L., 2009b, http://dx.doi.org/10.1111/j.1365-2966.2009.15331.x Monthly Notices of the Royal Astronomical Society, 399, 574

Wilkins S. M., Trentham N., Hopkins A. M., 2008, http://dx.doi.org/10.1111/j.1365-2966.2008.12885.x Monthly Notices of the Royal Astronomical Society, 385, 687

Wilkins S. M., Gonzalez-Perez V., Baugh C. M., Lacey C. G., Zuntz J., 2013a, http://dx.doi.org/10.1093/mnras/stt192 Monthly Notices of the Royal Astronomical Society, 431, 430

Wilkins S. M., et al., 2013b, http://dx.doi.org/10.1093/mnras/stt1471 Monthly Notices of the Royal Astronomical Society, 435, 2885

Wilkins S. M., Feng Y., Di-Matteo T., Croft R., Stanway E. R., Bunker A., Waters D., Lovell C., 2016, arXiv:1605.05044 [astro-ph]

Wilkins S. M., et al., 2019, arXiv e-prints, 1904, arXiv:1904.07504

Willott C. J., Percival W. J., McLure R. J., Crampton D., Hutchings J. B., Jarvis M. J., Sawicki M., Simard L., 2005, http://dx.doi.org/10.1086/430168 ApJ, 626, 657

Wise J. H., Demchenko V. G., Halicek M. T., Norman M. L., Turk M. J., Abel T., Smith B. D., 2014, http://dx.doi.org/10.1093/mnras/stu979 Monthly Notices of the Royal Astronomical Society, 442, 2560

Worthey G., 1994, http://dx.doi.org/10.1086/192096 The Astrophysical Journal Supplement Series, 95, 107

Wu H.-Y., Hahn O., Wechsler R. H., Mao Y.-Y., Behroozi P. S., 2013, http://dx.doi.org/10.1088/0004-637X/763/2/70 ApJ, 763, 70

Wylezalek D., et al., 2013, http://dx.doi.org/10.1088/0004-637X/769/1/79 ApJ, 769, 79

Yamada T., Nakamura Y., Matsuda Y., Hayashino T., Yamauchi R., Morimoto N., Kousai K., Umemura M., 2012, http://dx.doi.org/10.1088/0004-6256/143/4/79 AJ, 143, 79

Yu H., Wang F. Y., 2016, http://dx.doi.org/10.3847/0004-637X/820/2/114 ApJ, 820, 114

Yung L. Y. A., Somerville R. S., Finkelstein S. L., Popping G., Davé R., 2018, arXiv:1803.09761 [astro-ph]

Zackrisson E., et al., 2019, arXiv e-prints, 1905, arXiv:1905.00437

Zahid H. J., Dima G. I., Kewley L. J., Erb D. K., Davé R., 2012, http://dx.doi.org/10.1088/0004-637X/757/1/54 The Astrophysical Journal, 757, 54

Zahid H. J., Dima G. I., Kudritzki R.-P., Kewley L. J., Geller M. J., Hwang H. S., Silverman J. D., Kashino D., 2014, http://dx.doi.org/10.1088/0004-637X/791/2/130 The Astrophysical Journal, 791, 130

Zaroubi S., 2012, http://dx.doi.org/] arXiv:1206.0267 [astro-ph 10.1007/978-3-642-32362-1_2

Zhang Z.-Y., Romano D., Ivison R. J., Papadopoulos P. P., Matteucci F., 2018, http://dx.doi.org/10.1038/s41586-018-0196-x Nature, 558, 260